Exploring Mamba-based Diffusion: Towards better Rotation Invariance and Scan Aggregation

Vincent Tao Hu Felix Krause Ming Gui
Kim-Louis Simmoteit Johannes S. Fischer Björn Ommer
CompVis @ LMU Munich, MCML
https://taohu.me/zigma2

Abstract

This paper focuses on model design in Mamba-based generative modeling. State-space models like Mamba have gained popularity in the vision field for dense-pixel prediction tasks, thanks to their superior long sequence modeling capabilities. However, their training efficiency is often hampered by the inner-state summary and aggregation mechanism. Firstly, this paper identifies a previously overlooked issue in Mamba-based models: rotation invariance, which can benefit medical imaging and remote sensing. To address this, we suggest assessing Mamba-based generative models using generative metrics derived from a Dinov2-based backbone. Secondly, the scan path of visual image tokens plays a crucial role in this context of order-sensitive Mamba structure. Therefore, we propose a scan aggregation mechanism that combines the benefits of multi-scan per layer and layerwise scan heterogeneity. This results in a simple solution that only requires unidimensionally ordered tokens. This streamlined design enables us to apply any Mamba design improvements from language modeling to the vision field. Finally, we carried out comprehensive experiments to verify our method on various benchmarks. Our approach outperforms transformer-based and Mamba-based baselines in CelebA256, FaceHQ1024, and MS COCO256 across different resolutions and conditional scenarios. Our code will be released at https://taohu.me/zigma2.

1 Introduction

This paper discusses the design of State Space Models in generative models. State Space Models Gu et al. (2022b); Gupta et al. (2022); Gu et al. (2022a) have been widely explored in languages due to their efficiency in long sequence modeling Gu and Dao (2023) and comparative performance in in-context learning Park et al. (2024); Grazzi et al. (2024). Among these, Mamba Gu and Dao (2023) significantly enhanced the flexibility of SSMs in Language Modelling. It achieves this by relaxing the time-invariance constraint on SSM parameters, while still maintaining computational efficiency. Recently, Mamba has been generalized from unidimensional data to visual data such as images Hu et al. (2024); Zhu et al. (2024a); Liu et al.

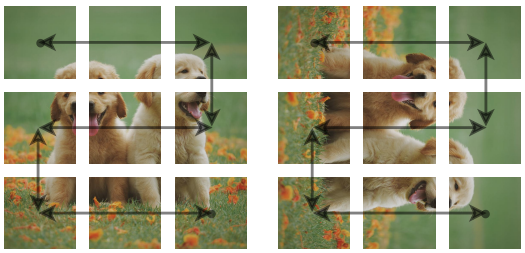

Figure 1: **Motivation: Bidirectional Scan Path in Mamba does not cover the case of rotated images.**

Preprint. Under review.

Table 1: **Comparison between different Mamba-based methods for vision modeling.**

Method	#Scan Path	Spatial-Continuity	Multi-Scan Per Layer	Layerwise Scan Heterogeneity
VisionMamba Zhu et al. (2024a)	2	✗	✓	✗
VMamba Liu et al. (2024c)	4	✗	✓	✗
Zigma Hu et al. (2024)	8	✓	✗	✓
Our	8	✓	✓	✓

(2024c); Yang et al. (2024a), videos Hu et al. (2024); Li et al. (2024b), point clouds Liang et al. (2024), human motion Zhang et al. (2024); Wang et al. (2024a), and even graphs Behrouz and Hashemi (2024). Furthermore, several studies indicate that Mamba's primary strength lies in its capability for long sequence modeling in the vision domain Yu and Bryant (2024); Zhu et al. (2024a).

In most applications of mamba-based models for 2D visual generation tasks, the scan path is a crucial concept. It typically describes how to traverse the image tokens that have been divided into patches, as shown in Figure 1. This process helps to better incorporate the inductive bias of vision data, especially given that the Mamba block accepts only unidimensional tokens by design. Therefore, based on the fact that they are derived from order-sensitive Mamba, the type of scan path (the order of input tokens) can be critical Hu et al. (2024). This differs from Transformer Vaswani et al. (2017) or Convolution-based He et al. (2016) networks, which are more permutation-invariant regard to performance and token orders. Although rotation augmentation in the dataset can empirically alleviate this issue, we demonstrate that it's still insufficient.

As depicted in Figure 1, using the default bidirectional scan in Mamba does not result in a rotation-invariant model. This can pose a significant problem in fields like medical imaging Xing et al. (2024) and remote sensing Zhao et al. (2024), where images inherently lack a concept of orientation. However, most current evaluation metrics such as Fréchet inception distance (FID) Heusel et al. (2017), Inception Score (IS) Salimans et al. (2016), and Kernel Inception Distance (KID) Bińkowski et al. (2018) are based on the InceptionV3 Szegedy et al. (2016) backbone. This backbone is pretrained on ImageNet and due to it being a convolution-based network without any inherent rotational invariance and with no focus on a learned invariance during training leading to inaccurate estimates when utilized as an evaluation metric in orientation-agnostic tasks. In this paper, we analyze this metric and propose evaluating the Mamba-based generative methods using the Dinov2 Oquab et al. (2023)-based metric, which includes Fréchet Distance and Kernel Distance.

To address rotation-related scenarios in model design, we propose a Scan Aggregation mechanism. This method involves implementing a range of scan paths across multiple identical image tokens for each layer. As shown in Table 1, this strategy combines the benefits of previous methods, which either stacked multiple scan paths in each layer Liu et al. (2024c); Zhu et al. (2024a), or applied various heterogeneous and spatial-continuous scan paths in each layer Hu et al. (2024).

Contributions. In summary, our contributions include:

1 Our analysis reveals that Mamba's current scan path is not suitable for standard datasets, as they require upright images. To address this, we propose a new benchmarking criterion that previous state-space models neglected when adapting the Mamba-based method from unidimensional data to 2D vision generative tasks.

2 We provide a simple method that utilizes a range of scan paths across multiple identical image tokens for each layer to make the network invariant to the rotation of the images.

3 We have carried out extensive experiments to validate our method on various benchmarks. Our method outperforms transformer-based and Mamba-based baselines in CelebA256, and FaceHQ1024, MS COCO256 across different resolutions and conditional scenarios.

2 Method

In this section, we begin with a brief introduction to the fundamentals of State-Space Models. We then explain why we chose to use rotation-augmented evaluation for Mamba-based backbone design. Following this, we detail our method of Scan Aggregation, which enhances rotation-awareness. Finally, we discuss the fundamentals of our diffusion theory, which is based on Stochastic Interpolants.

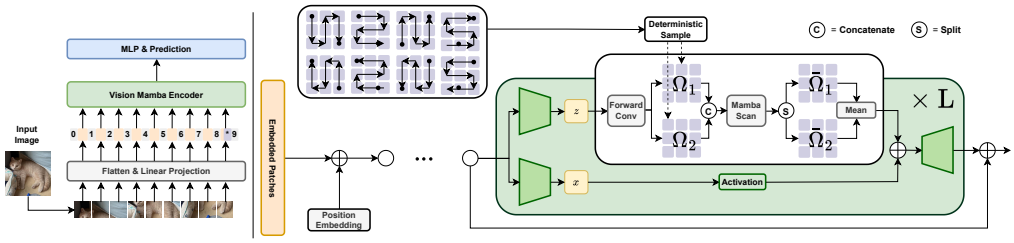

Figure 2: **Our Method.** We utilize a range of deterministically sampled scan paths across multiple identical image tokens for each layer. These are then concatenated along the token dimension and fed into the Mamba Scan block. Later, the concatenated image tokens are split and rearranged back to their original order, then averaged along the token dimension.

2.1 Preliminaries

State Space Models. State Space Models (SSMs) are typically used to model a continuous linear time-invariant (LTI) system where an input signal $z(t) \in \mathbb{R}$ is mapped to its output signal $y(t) \in \mathbb{R}$ through a state variable $h(t) \in \mathbb{R}^m$ with the following rules:

$$h'(t) = \mathbf{A}h(t) + \mathbf{B}z(t), \quad y(t) = \mathbf{C}h'(t) + \mathbf{D}z(t), \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{m\times m}$, $\mathbf{B} \in \mathbb{R}^{m\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times m}$ and $\mathbf{D} \in \mathbb{R}^{1\times 1}$ are parameters.

To make the above system usable for a discrete system, e.g. a sequence-to-sequence task, a timescale parameter $\mathbf{\Delta}$ is used to transform the parameters $\mathbf{A}$ and $\mathbf{B}$ to their discretized counterparts $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. In Mamba Gu and Dao (2023) and its following works Liu et al. (2024c); Zhu et al. (2024a), this is achieved with the following zero-order hold (ZOH) rule:

$$\bar{\mathbf{A}} = \exp(\mathbf{\Delta A}), \quad \bar{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}. \tag{2}$$

Afterwards, an input sequence $\{z_i\}$ (where $i = 1, 2, ...$) can be similarly mapped to its corresponding output sequence $\{y_i\}$:

$$h'_i = \bar{\mathbf{A}}h_{i-1} + \bar{\mathbf{B}}z_i, \quad y_i = \mathbf{C}h'_i + \mathbf{D}z_i. \tag{3}$$

Mamba. Since SSMs are often used to model LTI systems, their model parameters are shared by all time steps i. As identified in Mamba Gu and Dao (2023), the time-invariant nature of these parameters can significantly restrict the model's representational capabilities. To counter this, Mamba removes the time-invariant limitation, allowing the parameters $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{\Delta}$ to be dependent on the input sequence $\{x_i\}$. This process, referred to as the *selective scan*, leads to the token-dependent parameters $\{\mathbf{B}_i\}$, $\{\mathbf{C}_i\}$, and $\{\mathbf{\Delta}_i\}$. Specifically, the output sequence $\{y_i\}$ is computed from the $\{x_i\}$ as the following :

$$z_i = \sigma\big(\mathrm{DWConv}\big(\mathrm{Linear}(x_i)\big)\big), \quad x'_i = \sigma\big(\mathrm{Linear}(x_i)\big) \tag{4}$$

$$\mathbf{B}_i, \mathbf{C}_i, \mathbf{\Delta}_i = \mathrm{Linear}(z_i), \quad \bar{\mathbf{A}}_i, \bar{\mathbf{B}}_i = \mathrm{ZOH}(\mathbf{A}, \mathbf{B_i}, \mathbf{\Delta_i}) \tag{5}$$

$$h'_i = \bar{\mathbf{A}}_i h_{i-1} + \bar{\mathbf{B}}_i z_i, \quad y'_i = \mathbf{C}_i h'_i + \mathbf{D}z_i, \quad y_i = y'_i \odot x'_i, \tag{6}$$

where DWConv means the depthwise convolution operation, σ denotes the SiLU activation, and $\odot$ denotes element-wise multiplication.

2.2 Mamba-based Diffusion Backbone

tao: TODO Behrouz et al. (2024)

Rotation-augmented Evaluation for Mamba-based Backbone. Most of the previous studies Liu et al. (2024c); Zhu et al. (2024a) overlook the fact that the scan path is strictly and heuristically tied to the orientation of the images, as shown in Figure 1. This limitation hinders the smooth translation of progress when considering the rotation of images in the exploration of Mamba in vision tasks. This consideration could be potentially useful in medical imaging Xing et al. (2024) and

remote sensing Zhao et al. (2024). To compensate for this problem, we propose to evaluate the mamba-based generative methods on a rotation-augmented dataset using the Dinov2-based generative metrics, which we will motivate in the experimental part. To address this problem in the scenarios of rotations, we propose a mechanism of Scan Aggregation that can better incorporate the rotation-based evaluation.

Scan Aggregation. Previous studies Wang et al. (2022); Yan et al. (2023) have utilized bidirectional scanning within the structure of the SSM framework. This technique has been extended to encompass additional scanning directions Liu et al. (2024a,c); Yang et al. (2024g) to accommodate the unique characteristics of 2D image data. These methodologies unfold image patches in four different directions, thereby creating four unique sequences. Each of these sequences is then concurrently processed through every SSM. However, as each direction might have different SSM parameters ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$), an increase in the number of directions could potentially result in memory issues.

Our approach centers around the concept of token rearrangement before feeding them into the Mamba Scan block. For a given input feature $\mathbf{z}(k) \in \mathbb{R}^{T\times C}$ from layer k, the output feature $\mathbf{z}(k+1)$ of the Mamba Scan block after the rearrangement can be expressed as:

$$\mathbf{z}_{\Omega_k}(k) = \text{arrange}(\mathbf{z}(k), \Omega_k), \tag{7}$$

$$\bar{\mathbf{z}}_{\Omega_k}(k) = \text{scan}(\mathbf{z}_{\Omega_k}(k)), \tag{8}$$

$$\mathbf{z}(k+1) = \text{arrange}(\bar{\mathbf{z}}_{\Omega_k}(k), \bar{\Omega}_k), \tag{9}$$

where Ω_k represents the 1D permutation of layer k, which rearranges the order of the patch tokens by Ω_k,

$$\Omega_k(\cdot) = \text{map}[\Omega_k^0(\cdot), \Omega_k^1(\cdot), ..\Omega_k^n(\cdot)], \tag{10}$$

$$\bar{\Omega}_k(\cdot) = \text{reduce}[\bar{\Omega}_k^0(\cdot), \bar{\Omega}_k^1(\cdot), ..\bar{\Omega}_k^n(\cdot)], \tag{11}$$

where Ω_k^i and $\bar{\Omega}_k^i$ represent the reverse operation on the $\mathbb{R}^{T\times C}$. $\bar{\Omega}_k(\cdot), \bar{\Omega}_k(\cdot)$ is the overall mutual reverse operation on $\mathbb{R}^{nT\times C}$ to map the tensor into n various tensors and reduce them back. map denotes the copy operation and reduce denotes the mean operation among various scan paths and n is the scan path number per layer. Typically it can be deterministically sampled from m=8 possibilities following Hu et al. (2024). The operation of $\Omega_k(\cdot)$ ensures that both $\mathbf{z}(k)$ and $\mathbf{z}(k+1)$ maintain the sample order of the original image tokens.

Ultimately, the input size of the Mamba Scan is $\mathbb{R}^{nT\times C}$. This pattern simplifies the translation of the progress from the language modeling community of Vanilla Mamba Gu and Dao (2023) to the vision field. This is a stark contrast to previous vision-based methods Zhu et al. (2024a); Liu et al. (2024c), which required duplicating the state of SSM.

2.3 Diffusion Framework: Stochastic Interpolants

Previously, we discussed the model backbone design which enhances the incorporation of inductive bias from rotation-invariance. Next, we will introduce the Diffusion theory that guides our training and sampling processes.

Sampling based on vector $\mathbf{v}$ and score $\mathbf{s}$. Following Albergo et al. (2023); Tong et al. (2023), the time-dependent probability distribution $p_t(\mathbf{x})$ of $\mathbf{x}_t$ also coincides with the distribution of the reverse-time SDE Anderson (1982):

$$d\mathbf{X}_t = \mathbf{v}(\mathbf{X}_t, t)dt + \frac{1}{2}w_t\mathbf{s}(\mathbf{X}_t, t)dt + \sqrt{w_t}d\bar{\mathbf{W}}_t, \tag{12}$$

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

$$\begin{aligned}\mathbf{v}(\mathbf{x}, t) &= \mathbb{E}[\dot{\mathbf{x}}_t|\mathbf{x}_t = \mathbf{x}], \\ &= \dot{\alpha}_t\mathbb{E}[\mathbf{x}_*|\mathbf{x}_t = \mathbf{x}] + \dot{\sigma}_t\mathbb{E}[\boldsymbol{\varepsilon}|\mathbf{x}_t = \mathbf{x}],\end{aligned} \tag{13}$$

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 Song et al. (2021) or the reverse-time SDE (12).

Solving the reverse SDE (12) 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 Albergo et al. (2023), 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

$$\begin{aligned} \mathbf{x} &= \mathbb{E}[\mathbf{x}_t | \mathbf{x}_t = \mathbf{x}], \\ &= \alpha_t \mathbb{E}[\mathbf{x}_* | \mathbf{x}_t = \mathbf{x}] + \sigma_t \mathbb{E}[\boldsymbol{\varepsilon} | \mathbf{x}_t = \mathbf{x}], \end{aligned} \tag{14}$$

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

$$\mathbf{s}(\mathbf{x}, t) = \sigma_t^{-1} \frac{\alpha_t \mathbf{v}(\mathbf{x}, t) - \dot{\alpha}_t \mathbf{x}}{\dot{\alpha}_t \sigma_t - \alpha_t \dot{\sigma}_t}. \tag{15}$$

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 $\mathbf{s}$ and the velocity $\mathbf{v}$. It has been shown in score-based diffusion models Song et al. (2021) that the score can be estimated parametrically as $\mathbf{s}_\theta(\mathbf{x}, t)$ using the loss

$$\mathcal{L}_{\mathrm{s}}(\theta) = \int_0^T \mathbb{E}[\|\sigma_t \mathbf{s}_\theta(\mathbf{x}_t, t) + \boldsymbol{\varepsilon}\|^2] \mathrm{d}t. \tag{16}$$

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

$$\mathcal{L}_{\mathrm{v}}(\theta) = \int_0^T \mathbb{E}[\|\mathbf{v}_\theta(\mathbf{x}_t, t) - \dot{\alpha}_t \mathbf{x}_* - \dot{\sigma}_t \boldsymbol{\varepsilon}\|^2] \mathrm{d}t, \tag{17}$$

where θ represents the Mamba-based network with Scan Aggregation from the previous section. We adopt the linear path for training, due to its simplicity and relatively straight trajectory:

$$\alpha_t = 1 - t, \qquad \sigma_t = t. \tag{18}$$

We note that any time-dependent weight can be included under the integrals in both (16) and (17). These weight factors play a crucial role in score-based models when T becomes large Kingma and Gao (2023); Kingma et al. (2021). Thus, they provide a general form that considers both the time-dependent weight and the stochasticity.

3 Experiment

In this section, we first present the dataset, evaluation metrics and the specifics of our training process. We then justify our choices regarding evaluation and method design through a preliminary ablation study. Finally, we share our results from various datasets of differing resolutions, including Unconditional CelebA 256, COCO 256×256 and FFHQ 1024×1024.

3.1 Experimental Details

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 Xia et al. (2021) and FFHQ Karras et al. (2019), as employed in previous work such as Esser et al. (2021); Fischer et al. (2023); Hu et al. (2024). For ablation study about our method, we conduct it on MultiModal-CelebA 256×256 Xia et al. (2021) dataset without conditioning(text prompt), we conduct in single A100 GPU with batch size 16.

For text-conditioned generation, we conduct the experiments on MS COCO 256×256 Lin et al. (2014) datasets. Both datasets are composed of text-image pairs for training. Typically, there are 5 to 10 captions per image in MS COCO and MultiModal-CelebA. We convert discrete texts to a sequence of embeddings using a CLIP text encoder Radford et al. (2021) following Stable Diffusion Rombach et al. (2022). Then these embeddings are fed into the network as a sequence of tokens. For text-conditioning, we follow the pipeline from ZigMa Hu et al. (2024), the network structure can be found in the Appendix.

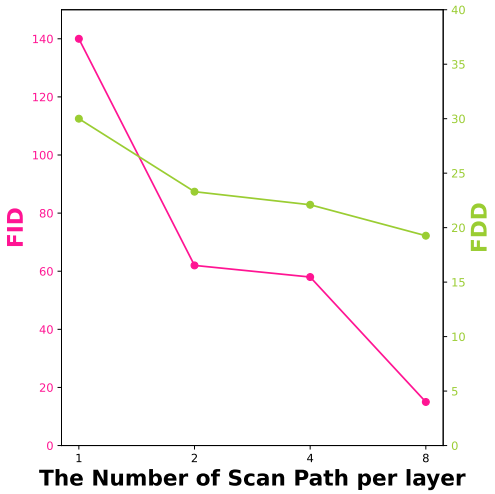

Figure 3: **FDD variation is steadier than FID with increased Scan Path in the rotation-augmented CelebA.**

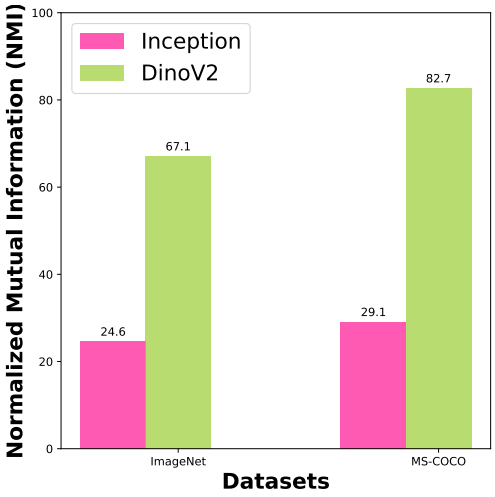

Figure 4: **Dino features exhibit greater robustness to the rotation when compared to Inception.**

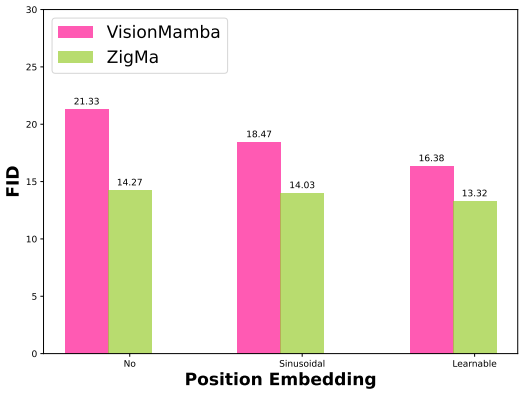

Figure 5: **Ablation of Position Embedding.**

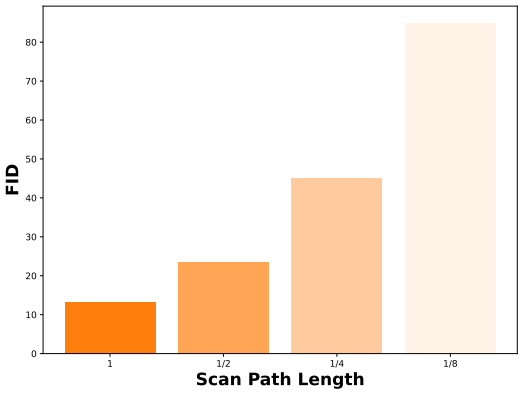

Figure 6: **Ablation of Length of the Scan Path.**

Rotation-Augmented Dataset. We create a new dataset that takes into account the object angle by performing a simple rotation of 0, $\pi/4$, $\pi/2$, and 1 radians. This rotation is applied to both training and evaluation images. Therefore, the metric based on Dinov2 Oquab et al. (2023) can reflect the capabilities of the mamba-based generative models.

Evaluation Metrics: FDD, KDD. As mentioned earlier, we primarily use the Dinov2-based feature for evaluation, leading to metrics such as Fréchet Dinov2 Distance (FDD) and Kernel Dinov2 Distance (KDD). In the preliminary ablation study experiments we largely report the FID as the default metric.

Training Details. We uniformly use AdamW Loshchilov and Hutter (2019) optimizer with a $1e-4$ learning rate. For extracting latent features, we employ an off-the-shelf VAE encoder. For sampling, we adopt the ODE sampling for speed consideration. For further details about training steps and GPU settings please refer to the Appendix A.2.

3.2 Experimental Results

3.2.1 Ablation study on Position Embedding

We primarily examine three types of Position Embedding: none, sinusoidal, and learnable. We focus on two previous works, VisionMamba Zhu et al. (2024a) and ZigMa Hu et al. (2024). VisionMamba

Table 2: **Ablation about our method on CelebA 256×256 dataset**. $\underline{n}\text{In}\underline{m}$ denotes that we deploy $\underline{n}$ scan paths per layer, with a total of $\underline{m}$ scan paths.

Method		State Dim	Unrotated ↓		Rotated ↓		#Param
			FDD ↓	KDD ↓	FDD ↓	KDD ↓	
DiT Peebles and Xie (2023)		-	27.75	41.94	32.29	28.13	256M
VisionMamba Zhu et al. (2024a)		16	31.80	42.12	36.48	55.52	140M
Zigma Hu et al. (2024)		16	26.76	34.21	32.02	50.73	135M
Our	2In4	16	23.38	20.34	30.25	46.68	135M
	4In8	16	35.50	45.10	41.20	61.20	136M
	2In8	32	23.21	30.23	28.89	45.29	136M
VisionMamba w/ *Our*		16	29.25	39.44	35.85	56.53	134M
Our	2In8	16	**21.93**	**28.47**	**27.16**	**41.56**	134M

considers two scan paths (forward and backward) without spatial continuity, while ZigMa considers eight scan paths with spatial continuity.

From Figure 5, we find that both multiple scan paths and spatial continuity are essential for satisfying performance. Additionally, learnable Position Embedding can further extract the potential of Mamba-based generative models, even with eight spatially continuous scan paths. This finding encourages us to pursue broader scan aggregation in our next experiment.

3.2.2 Why Dinov2 feature is better than Inception for Rotation-augmented dataset?

Most metrics for generative models are based on the 1D pooled feature from the backbone, so we evaluate the discriminative ability of these features. Specifically, we apply a k-means clustering with a fixed number and compare the Normalized Mutual Information (NMI) between the representation of unrotated and rotated images. This comparison should reflect the discriminative ability of our method.

We conduct experiments mainly on two datasets: the object-centric ImageNet Deng et al. (2009) and the non-object-centric MS COCO Lin et al. (2014). As shown in Figure 4, the DinoV2-based backbone Oquab et al. (2023) demonstrates better robustness compared to the InceptionV3 backbone Szegedy et al. (2016). Therefore, we recommend using this backbone for evaluating generative models on rotation-augmented datasets. Our hypothesis is that other unsupervised backbones can also be utilized in this pipeline, but we chose DinoV2 for simplicity.

3.2.3 Preliminary Ablation of Scan Path and Scan Length

In this preliminary ablation study, we primarily investigate the impact of the number and length of scan paths in generative models. In Figure 3, we demonstrate that as we incrementally include more types of scan paths, we consistently see improvements from the Mamba-based structure, which motivates us to design our scan aggregation method.

In Figure 6, we observe that decreasing the scan path length from 1 to 1/2, 1/4, 1/8 drastically reduces performance, underscoring the importance of long scan lengths. This outcome is expected, because with each layer the network can only account for a portion of the image patch tokens and completely ignores the remaining tokens, which, in turn, negatively impacts overall performance.

3.2.4 Main Ablation

In this part, we mainly ablate various elements of our method, including the scan aggregation manner, and state dimension. We mainly compare with the following baselines: a transformer-based backbone: DiT Peebles and Xie (2023), VisionMamba Zhu et al. (2024a) and ZigMa Hu et al. (2024). We primarily aim to compare with them in two different scenarios: the unrotated version and the rotated version. Based on previous observations, we primarily use two metrics based on DinoV2 Oquab et al. (2023) for evaluation: Fréchet Distance and Kernel Distance. To ensure a fair comparison, we adjust the number of layers and block dimensions to make the parameter count nearly identical.

Table 3: **Text-conditioned image generation on MS COCO 256×256 dataset.**

Method	Unrotated ↓		Rotated ↓		#Param
	FDD ↓	KDD ↓	FDD ↓	KDD ↓	
DiT Peebles and Xie (2023)	49.10	65.40	47.91	54.01	133M
VisionMamba Zhu et al. (2024a)	45.10	55.69	47.08	56.55	140M
ZigMa Hu et al. (2024)	57.26	73.65	57.16	71.21	134M
Our	**43.16**	**53.49**	**43.47**	**52.22**	136M

Table 4: **Unconditional image generation of FacesHQ 1024×1024 dataset. OOM denotes Out of Memory due to GPU memory limitation.**

Method	Unrotated ↓		Rotated ↓		#Param
	FDD ↓	KDD ↓	FDD ↓	KDD ↓	
DiT Peebles and Xie (2023)	OOM	OOM	OOM	OOM	-
ZigMa Hu et al. (2024)	79.10	45.20	82.10	160.1	137M
Our	**74.30**	**40.10**	**76.90**	**142.0**	137M

As shown in Table 2, we make the following observations. 1). Compared to VisionMamba Zhu et al. (2024a), which only replicates the state for forward and backward scans without considering spatial continuity, most of our variants consistently outperform it. This underscores the effectiveness of our method. 2). If the scan path per layer is too large, such as 4 scan paths, it can negatively affect the result. 3). Further increasing the state of the SSM can enhance the result, even when the token number is doubled from the same image but via different scan paths. 4). We explore to apply our method to VisionMamba. This involves concatenating the forward and backward scans into a single tensor of $\mathbb{R}^{2T \times C}$, which is then fed into the Mamba block. By applying our method to VisionMamba, we observe some improvements. In our remaining experiments, we will keep using the optimal setting of 2In8.

3.2.5 Result on Text-conditioned image generation on MS COCO 256×256 dataset

We explore the method within the framework of text-conditioned generation, by conditioning the text prompt feature with a cross-attention block added to the Mamba block, as outlined in Hu et al. (2024). More details can be found in the appendix. As shown in Table 3, our method outperforms related baselines, indicating that it is effective in both rotated and unrotated scenarios.

3.2.6 Result on high-resolution FacesHQ 1024×1024 dataset

We explore this in a high-resolution dataset, primarily comparing it with the Transformer-based backbone DiT Peebles and Xie (2023) and a typical Mamba-based backbone Zigma Hu et al. (2024). As shown in Table 4, our method demonstrates superior performance in terms of FDD and KDD compared to the baseline, given the same number of parameters.

We visualized the generated images of various resolutions in Figure 7. Despite some noticeable artifacts, the results are generally satisfactory. We believe that these artifacts could be eliminated by using more computing resources.

4 Related Works

4.1 Generalizing State Space Models from Unidimensional Data to 2D Visual Data

The State Space Models Gu et al. (2021, 2022b); Gu and Dao (2023), initially applied in language modeling, has recently been generalized to visual tasks due to its advantage in long-context modeling. For visual perception tasks, CNNs He et al. (2016); Huang et al. (2017); Krizhevsky et al. (2012) and ViTs Dosovitskiy et al. (2021) have been the most commonly used architectures for a long time. An early application of SSMs into the field of vision is S4Nd which extends the kernel from 1D to 2D using an outer-product Nguyen et al. (2022). However, methods like Mamba and RKWV have had great success in general language modeling Gu and Dao (2023); Peng et al. (2023). Naturally,

Figure 7: Visualization of the FacesHQ 1024×1024 dataset.

following this, these architectures have been applied to a variety of vision tasks. VisionMamba utilizes a bidirectional model for classification tasks. Liu et al. (2024c) propose a Visual State Space model, VMamba by introducing scanning paths that divide the inherently non-sequential image into sequential image patches that can then be processed by Mamba. There are other similar works motivated by linear attention like RWKV Peng et al. (2023, 2024); Duan et al. (2024); Fei et al. (2024b). For more generalized State Space Models in vision data, we refer the reader to the survey papers Xu et al. (2024); Liu et al. (2024b).

Most research aims to solve the problem by considering the scan path in both directions Zhu et al. (2024a); Fei et al. (2024a); Li et al. (2024c); Xing et al. (2024); Wang et al. (2024b); Yang et al. (2024g); Wu et al. (2024a). On the other hand, it's also important to aggregate various scan directions in each layer. This can be typically approached in two ways. Firstly, by duplicating the token patches and the inner state of SSM in each neural network layer to reason different scan paths Zhu et al. (2024a); Xing et al. (2024), which can increase learning efficiency. Secondly, by applying various zigzag scan paths in different layers Hu et al. (2024); Yang et al. (2024a), which can enhance memory and parameter efficiency. As shown in Table 1, our work attempts to combine the best of both approaches, allowing us to incorporate multiple scan paths and increase the reasoning ability per layer.

4.2 State Space Models based Generative Models

State Space Models (SSMs) have recently gained significant interest in generative modeling. DiffSSM Yan et al. (2023) focuses on both unconditional and class-conditioned scenarios within the S4 framework Gu et al. (2022b). ZigMa Hu et al. (2024) proposes a Mamba-based DiT-style generative backbone for image synthesis and incorporates a selective scan pathing that takes spatial continuity for images into account. Reversely, DiS Fei et al. (2024a) explores the Mamba in generation by a UViT-based style Bao et al. (2023). The work by Ju et al. Ju and Zhou (2024) and Zheng et al. Zheng and Zhang (2024) focus on SSM-CNN hybrid architectures in the field of medical image synthesis and extending those with attention respectively. Gao et al. Gao et al. (2024) investigate the Mamba architecture for video synthesis, concluding that the attention mechanism is best used for capturing local spatiotemporal details, whereas Mamba excels at understanding global patterns. Furthermore, RWKV Peng et al. (2023) was proposed as a promising alternative to traditional transformers. Recent work by Fei et al. (2024b) further validates the scalability and efficacy of RWKV for image synthesis. Our work also utilizes the Mamba backbone. Our goal is to address the rotation invariance in generative modeling by introducing a novel Scan Aggregation mechanism.

5 Conclusion

This paper primarily focuses on the rotation invariance representation in mamba-based generative modeling. To address this, we propose evaluating the rotation-augmented dataset using a DinoV2-based backbone for generative evaluation. Further, we introduce a Scan Aggregation to format our image tokens as unidimensional data, which simplifies the translation of State Space Models into the vision domain. Ultimately, 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 **?**, GLA Yang et al. (2024c), and several others listed at FLA Yang and Zhang (2024)[1].

6 Acknowledgements

We would like to thank Timy Phan, Nick Stracke for the extensive proofreading. Thank Owen Vincent and Stefan Baumann for the technical support. The authors acknowledge the Gauss Center for Supercomputing for providing compute through the NIC on JUWELS at JSC. We would like to thank Munich Center for Machine Learning (MCML) for the financial funding. The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Climate Action within the project NXT GEN AI METHODS Generative Methoden für Perzeption, Prädiktion und Planung". The authors would like to thank the consortium for the successful cooperation.

7 Limitations

The primary limitation we faced was GPU resource constraints, which prevented us from exploring longer training durations. However, we expect similar outcomes if we had been able to do so. The second limitation involves the increased computation resulting from the doubling of the token number length in scan aggregation. Although the time complexity associated with token length is not as significant as in transformers, it's still a factor. Our method is simple and general. We aim to create a generative framework that can effectively bridge the language and vision communities. In future work, we plan to explore various applications of our method. By leveraging its scalability for long-sequence modeling, we hope to improve the utilization of the Mamba framework across different domains and applications.

8 Broader Impacts

This work aims to improve the scalability and potential of the Mamba algorithm within diffusion models, allowing for the generation of large, high-fidelity images. However, like other efforts to enhance large-scale image synthesis models, our approach carries the risk of generating harmful or deceptive content. Therefore, it's crucial to consider ethical implications and implement safeguards to mitigate these risks.

References

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

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

Archit, A. and Pape, C. (2024). Vim-unet: Vision mamba for biomedical segmentation. *arXiv preprint arXiv:2404.07705*.

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

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

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

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

Behrouz, A., Santacatterina, M., and Zabih, R. (2024). Mambamixer: Efficient selective state space models with dual token and channel selection.

Benjdira, B., Bazi, Y., Koubaa, A., and Ouni, K. (2019). Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*, 11(11):1369.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.

Cao, Q., Chen, Y., Ma, C., and Yang, X. (2023). Few-shot rotation-invariant aerial image semantic segmentation.

Cheng, G., Han, J., and Lu, X. (2017). Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):18651883.

Cheng, G., Zhou, P., and Han, J. (2016). Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 54(12):7405–7415.

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545563.

Cossio, M. (2023). Augmenting medical imaging: A comprehensive catalogue of 65 techniques for enhanced data analysis.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *CVPR*.

Dong, W., Zhu, H., Lin, S., Luo, X., Shen, Y., Liu, X., Zhang, J., Guo, G., and Zhang, B. (2024). Fusion-mamba for cross-modality object detection. *ArXiv*, abs/2404.09146.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Du, C., Li, Y., and Xu, C. (2024). Understanding robustness of visual state space models for image classification.

Duan, Y., Wang, W., Chen, Z., Zhu, X., Lu, L., Lu, T., Qiao, Y., Li, H., Dai, J., and Wang, W. (2024). Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. *arXiv*.

Esser, P., Rombach, R., and Ommer, B. (2021). Taming transformers for high-resolution image synthesis. In *CVPR*.

Fang, Z., Wang, Y., Wang, Z., Zhang, J., Ji, X., and Zhang, Y. (2024). Mammil: Multiple instance learning for whole slide images with state space models.

Fei, Z., Fan, M., Yu, C., and Huang, J. (2024a). Scalable diffusion models with state space backbone. *arXiv*.

Fei, Z., Fan, M., Yu, C., Li, D., and Huang, J. (2024b). Diffusion-rwkv: Scaling rwkv-like architectures for diffusion models. *arXiv*.

Fischer, J. S., Gui, M., Ma, P., Stracke, N., Baumann, S. A., and Ommer, B. (2023). Boosting latent diffusion with flow matching. *arXiv*.

Gao, Y., Huang, J., Sun, X., Jie, Z., Zhong, Y., and Ma, L. (2024). Matten: Video generation with mamba-attention. *arXiv*.

Gong, H., Kang, L., Wang, Y., Wan, X., and Li, H. (2024). nnmamba: 3d biomedical image segmentation, classification and landmark detection with state space model. *arXiv*.

Grazzi, R., Siems, J., Schrodi, S., Brox, T., and Hutter, F. (2024). Is mamba capable of in-context learning? *arXiv preprint arXiv:2402.03170*.

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

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

Gu, A., Goel, K., and Ré, C. (2022b). Efficiently modeling long sequences with structured state spaces. In *ICLR*.

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

Guo, T., Wang, Y., Shu, S., Chen, D., Tang, Z., Meng, C., and Bai, X. (2024). MambaMorph: a Mamba-based Framework for Medical MR-CT Deformable Registration. *arXiv e-prints*, page arXiv:2401.13934.

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

Hao, J., He, L., and Hung, K. F. (2024). T-mamba: Frequency-enhanced gated long-range dependency for tooth 3d cbct segmentation. *ArXiv*, abs/2404.01065.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hee, K., Cosa, A., Santhanam, N., Jannesari, M., Maros, M., and Ganslandt, T. (2022). Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*.

Hu, V. T., Baumann, S. A., Gui, M., Grebenkova, O., Ma, P., Fischer, J., and Ommer, B. (2024). Zigma: A dit-style zigzag mamba diffusion model. In *Arxiv*.

Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *CVPR*.

Huang, T., Pei, X., You, S., Wang, F., Qian, C., and Xu, C. (2024). Localmamba: Visual state space model with windowed selective scan. *arXiv preprint arXiv:2403.09338*.

Ju, Z. and Zhou, W. (2024). Vm-ddpm: Vision mamba diffusion for medical image synthesis. *arXiv*.

Kalra, A., Stoppi, G., Brown, B., Agarwal, R., and Kadambi, A. (2021). Towards rotation invariance in object detection.

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

Kingma, D., Salimans, T., Poole, B., and Ho, J. (2021). Variational diffusion models. In *NeurIPS*.

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

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS*.

Li, H., Hu, Q., Yao, Y., Yang, K., and Chen, P. (2024a). CFMW: Cross-modality Fusion Mamba for Multispectral Object Detection under Adverse Weather Conditions. *arXiv e-prints*, page arXiv:2404.16302.

Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., and Qiao, Y. (2024b). Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*.

Li, S., Singh, H., and Grover, A. (2024c). Mamba-nd: Selective state space modeling for multi-dimensional data. *arXiv*.

Liang, D., Zhou, X., Wang, X., Zhu, X., Xu, W., Zou, Z., Ye, X., and Bai, X. (2024). Pointmamba: A simple state space model for point cloud analysis. *arXiv*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

Linmans, J., Winkens, J., Veeling, B. S., Cohen, T. S., and Welling, M. (2018). Sample efficient semantic segmentation using rotation equivariant convolutional networks.

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

Liu, X., Zhang, C., and Zhang, L. (2024b). Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404*.

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

Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *ICLR*.

Lu, J., Yao, J., Zhang, J., Zhu, X., Xu, H., Gao, W., Xu, C., Xiang, T., and Zhang, L. (2021). Soft: softmax-free transformer with linear complexity. In *NeurIPS*.

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

Ma, X., Zhang, X., and Pun, M.-O. (2024b). Rs3mamba: Visual state space model for remote sensing images semantic segmentation.

Marmanis, D., Datcu, M., Esch, T., and Stilla, U. (2016). Deep learning earth observation classification using imagenet pretrained networks. *IEEE Geoscience and Remote Sensing Letters*, 13(1):105109.

Nguyen, E., Goel, K., Gu, A., Downs, G., Shah, P., Dao, T., Baccus, S., and Ré, C. (2022). S4nd: Modeling images and videos as multidimensional signals with state spaces. *NeurIPS*.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Park, J., Park, J., Xiong, Z., Lee, N., Cho, J., Oymak, S., Lee, K., and Papailiopoulos, D. (2024). Can mamba learn how to learn? a comparative study on in-context learning tasks. *ICML*.

Peebles, W. and Xie, S. (2023). Scalable diffusion models with transformers. *ICCV*.

Peng, B., Alcaide, E., Anthony, Q. G., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., Kranthikiran, G., He, X., Hou, H., Kazienko, P., Koco, J., Kong, J., Koptyra, B., Lau, H., Mantri, K. S. I., Mom, F., Saito, A., Tang, X., Wang, B., Wind, J. S., Wozniak, S., Zhang, R., Zhang, Z., Zhao, Q., Zhou, P., Zhu, J., and Zhu, R. (2023). Rwkv: Reinventing rnns for the transformer era. In *Conference on Empirical Methods in Natural Language Processing*.

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

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

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

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *NeurIPS*.

Sanjid, K. S., Hossain, M. T., Junayed, M. S. S., and Uddin, D. M. M. (2024). Integrating mamba sequence model and hierarchical upsampling network for accurate semantic segmentation of multiple sclerosis legion.

Shen, Q., Yi, X., Wu, Z., Zhou, P., Zhang, H., Yan, S., and Wang, X. (2024). Gamba: Marry gaussian splatting with mamba for single view 3d reconstruction. *ArXiv*, abs/2403.18795.

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

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

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *CVPR*.

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

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

Wang, J., Chen, J., Chen, D. Z., and Wu, J. (2024a). Large window-based mamba unet for medical image segmentation: Beyond convolution and self-attention. *ArXiv*, abs/2403.07332.

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

Wang, Q., Wang, C., Lai, Z., and Zhou, Y. (2024b). Insectmamba: Insect pest classification with state space model. *arXiv preprint arXiv:2404.03611*.

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

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

Wang, X., Huang, Z., Zhang, S., Zhu, J., and Feng, L. (2024). GMSR:Gradient-Guided Mamba for Spectral Reconstruction from RGB Images. *arXiv e-prints*, page arXiv:2405.07777.

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

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

Wu, R., Liu, Y., Liang, P., and Chang, Q. (2024a). H-vmunet: High-order vision mamba unet for medical image segmentation.

Wu, R., Liu, Y., Liang, P., and Chang, Q. (2024b). Ultralight vm-unet: Parallel vision mamba significantly reduces parameters for skin lesion segmentation. *arXiv preprint arXiv:2403.20035*.

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

Xie, J., Liao, R., Zhang, Z., Yi, S., Zhu, Y., and Luo, G. (2024). Promamba: Prompt-mamba for polyp segmentation.

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

Xu, R., Yang, S., Wang, Y., Du, B., and Chen, H. (2024). A survey on vision mamba: Models, applications and challenges. *arXiv preprint arXiv:2404.18861*.

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

Yang, C., Chen, Z., Espinosa, M., Ericsson, L., Wang, Z., Liu, J., and Crowley, E. J. (2024a). Plainmamba: Improving non-hierarchical mamba in visual recognition. *arXiv preprint arXiv:2403.17695*.

Yang, G., Du, K., Yang, Z., Du, Y., Zheng, Y., and Wang, S. (2024b). Cmvim: Contrastive masked vim autoencoder for 3d multi-modal representation learning for ad classification. *arXiv preprint arXiv:2403.16520*.

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

Yang, S., Wang, Y., and Chen, H. (2024d). Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. *arXiv preprint arXiv:2403.06800*.

Yang, S., Wang, Y., and Chen, H. (2024e). Mambamil: Enhancing long sequence modeling with sequence reordering in computational pathology. *ArXiv*, abs/2403.06800.

Yang, S. and Zhang, Y. (2024). Fla: A triton-based library for hardware-efficient implementations of linear attention mechanism.

Yang, Y., Ma, C., Yao, J., Zhong, Z., Zhang, Y., and Wang, Y. (2024f). Remamber: Referring image segmentation with mamba twister.

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

Yu, Weihao Wang, X. and Bryant, K. (2024). Mambaout: Do we really need mamba for vision? *arXiv preprint arXiv:2405.07992*.

Yue, Y. and Li, Z. (2024). Medmamba: Vision mamba for medical image classification.

Zhang, J., Zhang, Y., and Xu, X. (2021). Objectaug: Object-level data augmentation for semantic image segmentation.

Zhang, Z., Liu, A., Reid, I., Hartley, R., Zhuang, B., and Tang, H. (2024). Motion mamba: Efficient and long sequence motion generation with hierarchical and bidirectional selective ssm. *arXiv preprint arXiv:2403.07487*.

Zhao, S., Chen, H., Zhang, X., Xiao, P., Bai, L., and Ouyang, W. (2024). Rs-mamba for large remote sensing image dense prediction. *arXiv preprint arXiv:2404.02668*.

Zheng, Z. and Zhang, J. (2024). Fd-vision mamba for endoscopic exposure correction. *arXiv preprint arXiv:2402.06378*.

Zhou, H., Wu, X., Chen, H., Chen, X., and He, X. (2024). RSDehamba: Lightweight Vision Mamba for Remote Sensing Satellite Image Dehazing. *arXiv e-prints*, page arXiv:2405.10030.

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

Zhu, Q., Cai, Y., Fang, Y., Yang, Y., Chen, C., Fan, L., and Nguyen, A. (2024b). Samba: Semantic segmentation of remotely sensed images with state space model. *arXiv preprint arXiv:2404.01705*.

A Appendix

A.1 Additional related works

The recent success of Mamba as a backbone for vision tasks is visible by outperforming convolutional and transformer-based models in various vision subtasks such as classification Zhu et al. (2024a); Liu et al. (2024c); Li et al. (2024b); Gong et al. (2024); Li et al. (2024c); Yue and Li (2024); Fang et al. (2024); Huang et al. (2024); Du et al. (2024); Behrouz et al. (2024); Wang et al. (2024b); Huang et al. (2024), segmentation Xing et al. (2024); Ma et al. (2024a); Liu et al. (2024a); Wu et al. (2024a); Wang et al. (2024b); Zhu et al. (2024b); Gong et al. (2024); Xie et al. (2024); Wu et al. (2024a); Hao et al. (2024); Sanjid et al. (2024); Ma et al. (2024b); Yang et al. (2024f); Behrouz et al. (2024); Zhu et al. (2024b); Wu et al. (2024b); Archit and Pape (2024); Wang et al. (2024a), reconstruction Zhou et al. (2024); Wang et al. (2024); Shen et al. (2024), and much more Guo et al. (2024); Yang et al. (2024d,b). Similar to self-attention replacements cross modal Mamba blocks Dong et al. (2024); Yang et al. (2024e); Li et al. (2024a) have been designed for cross-attention replacement allowing to fuse information from multiple branches.

Generalizing SSM from unidimensional data to 2D Visual Data Transformer architectures Vaswani et al. (2017) have become a cornerstone for numerous state-of-the-art models, known for their high performance and ease of training due to their parallelizability. However, the attention mechanism in transformers scales quadratically with sequence length Lu et al. (2021), making computations on long sequences challenging. State-Space Models (SSMs) offer better scalability but have historically underperformed compared to transformers. Despite SSMs' universal approximation capabilities Wang and Xue (2024); Wang and Li (2023); Wang and Xue (2024), Mamba Gu and Dao (2023) is the first SSM to match transformer performance.

For visual perception tasks, CNNs He et al. (2016); Huang et al. (2017); Krizhevsky et al. (2012) and ViTs Dosovitskiy et al. (2021) have been the most commonly used architectures for a long time. An early application of SSMs into the field of vision is S4Nd which extends the kernel from 1D to 2D using an outer-product Nguyen et al. (2022). However, the already mentioned methods like Mamba and RKWV have had great success in general language modeling Gu and Dao (2023); Peng et al. (2023). Naturally, following this, these architectures have been applied to a variety of vision tasks. VisionMamba utilizes a bidirectional model for classification tasks. Liu et al. (2024c) propose a Visual State-Space model, VMamba by introducing scanning paths that divide the inherently non-sequential image into sequential image patches that can then be processed by Mamba. Vision RKWV (VRWKV) extends the original RWKV architecture to vision tasks by incorporating a Q-Shift that creates a localized flow of information while a Bi-WKV module is used as a global attention mechanism Duan et al. (2024).

Most research aims to solve the problem by considering the scan path in both directions Zhu et al. (2024a); Fei et al. (2024a). On the other hand, it's also important to stack various scan directions in each layer. This can be approached in two ways. Firstly, by duplicating the token patches and the inner state in each neural network layer to reason different scan paths Zhu et al. (2024a), which can increase the learning efficiency. Secondly, by applying various zigzag scan paths in different layers Hu et al. (2024); Yang et al. (2024a), which can enhance memory and parameter efficiency. This work attempts to combine the best of both approaches, allowing us to incorporate multiple scan paths and increase the reasoning ability per layer.

All of these studies overlook the fact that the scan path is strictly tied to the orientation of the images. This limitation hinders the smooth translation of progress when considering the rotation of images in the exploration of the Mamba in vision task. This consideration could be potentially useful in medical imaging Xing et al. (2024) and remote sensing Zhao et al. (2024).

State-space model based Generative Backbones State-Space Models (SSMs) have recently gained significant interest in generative modeling. DiffSSM Yan et al. (2023) focuses on both unconditional and class-conditioned scenarios within the S4 framework Gu et al. (2022b). DiS Fei et al. (2024a) delves into SSMs on a smaller scale. The work by Ju et al. Ju and Zhou (2024) and Zheng et al. Zheng and Zhang (2024) focus on SSM-CNN hybrid architectures in the field of medical image synthesis and extending those with attention respectively. Hu et al. (2024) proposes a Mamba-based generative backbone for image synthesis and incorporates a selective scan pathing that takes spatial continuity for images into account. Gao et al. Gao et al. (2024) investigate the Mamba architec-

Table 5: **Hyperparameters and number of parameters for our network in various datasets.**

	FacesHQ 1024	MS-COCO 256	MultiModal-CelebA 256
Autoencoder f	8	8	8
z-shape	$4 \times 128 \times 128$	$4 \times 32 \times 32$	$4 \times 32 \times 32$
Model size	133.8M	133.8M	133.8M
Patch size	2	1	2
Channels	768	768	768
Depth	24	24	24
Optimizer	AdamW	AdamW	AdamW
Batch size/GPU	4	16	16
GPU num	4	4	1
Learning rate	1e-4	1e-4	1e-4
weight decay	0	0	0
EMA rate	0.9999	0.9999	0.9999
Warmup steps	0	0	0
step number	50k	50k	50k
GPU A100	4	4	1

ture for video synthesis, concluding that the attention mechanism is best used for capturing local spatiotemporal details, whereas Mamba excels at understanding global patterns.

Furthermore, RWKV Peng et al. (2023) was proposed as a promising alternative to traditional transformers. Recent work by Fei et al. Fei et al. (2024b) further validates the scalability and efficacy of RWKV for image synthesis.

Different fields in computer vision have varying requirements. General tasks benefit from scale, while niche fields need to adapt state-of-the-art solutions on a smaller scale to their specific needs Benjdira et al. (2019); Hee et al. (2022); Marmanis et al. (2016); Cheng et al. (2017). For instance, object detection and segmentation tasks must be consistent across different rotations to provide reliable predictions Cheng et al. (2016); Kalra et al. (2021); Linmans et al. (2018); Zhang et al. (2021); Cao et al. (2023). Even entire fields like medical imaging and aerial imaging rely on specific characteristics and biases to convey information. Many modalities in the former do not have a set rotational orientation (list of modalities) while the latter completely renounces rotational orientation Cao et al. (2023); Chlap et al. (2021); Cossio (2023).

A.2 Details

A.2.1 Training Details

For DiT baseline, we set layer=28, dim=512 to make the parameter number approximately around 130M to make the comparison fair.

We illustrate the text-conditioned backbone structure in Figure 8.

We list the training details in Table 5. We by default use the state dimension of 16.

A.2.2 Evaluation Details

Implementation of the evaluation metric FDD,KDD. We implemented it based on the public repo `https://github.com/layer6ai-labs/dgm-eval` from paper Stein et al. (2023). For KDD, we display only the number obtained after multiplying by 100.

The experiment details about NMI calculation. We calculated the Normalized Mutual Information (NMI) score for the ImageNet Deng et al. (2009) validation dataset, which contains 50,000 images and the COCO validation dataset Lin et al. (2014) containing 80 object categories spread over 40,504 images. Then we extracted intermediate representations from InceptionV3 Szegedy et al. (2016) and DinoV2 Oquab et al. (2023).

We then applied k-means clustering to partition the feature space into 1,000 and 80 classes respectively. This process was performed on both the original validation set and a 90-degree rotated ver-

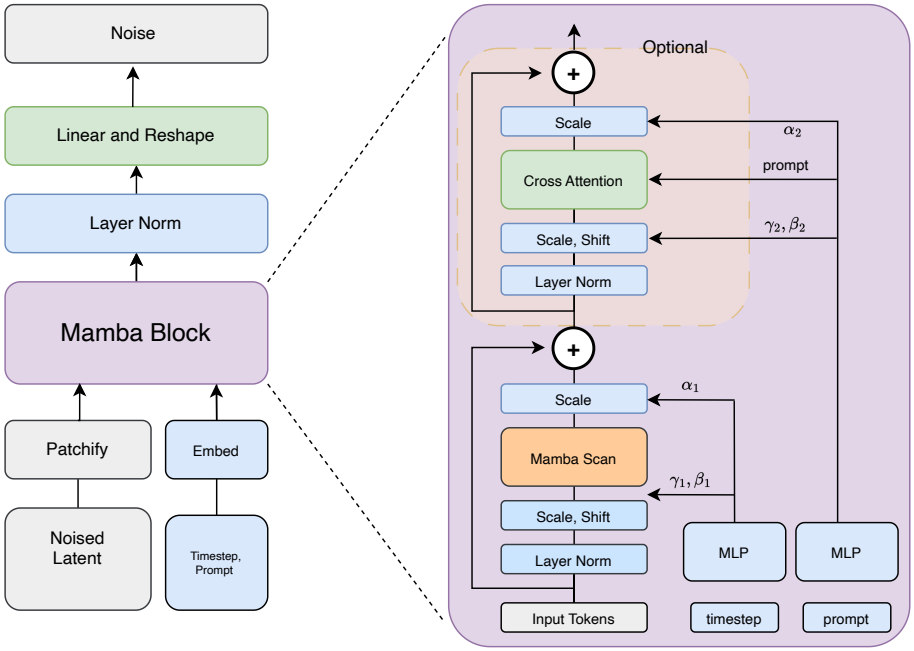

Figure 8: The Mamba based backbone design we use for text-conditioned generation. Images adapted from Hu et al. (2024).

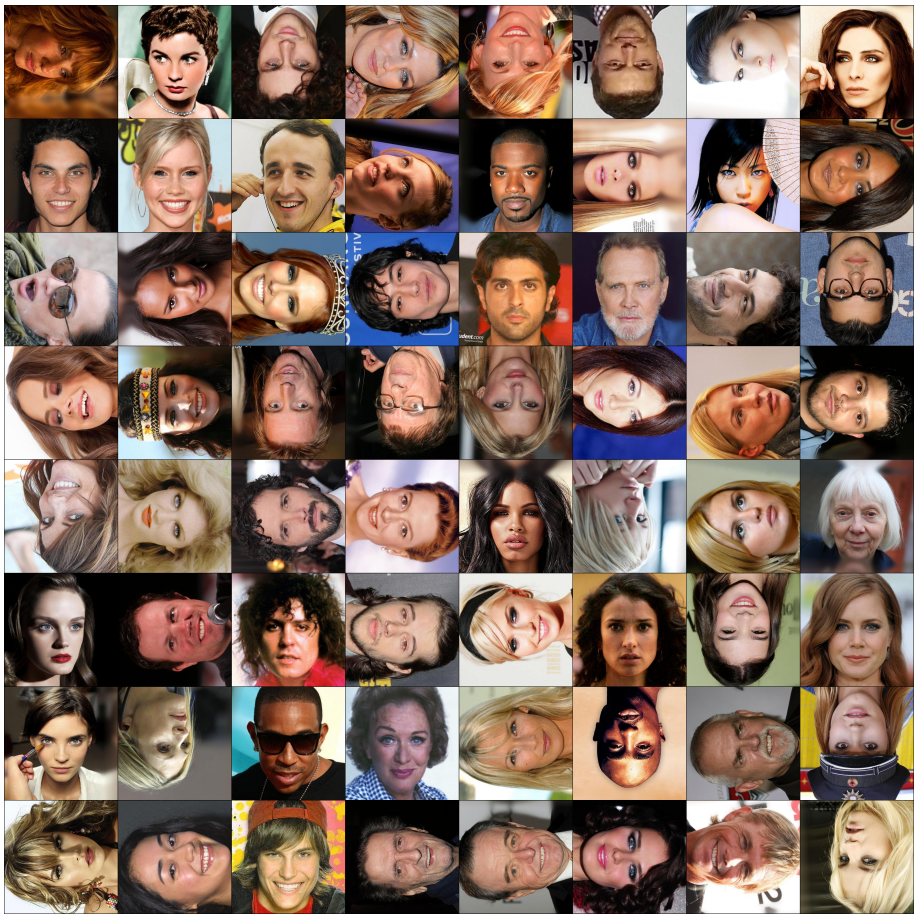

Figure 9: Visualization of rotated images.

sion of the validation set. The resulting cluster labels from both sets were used to calculate the NMI, showing how much the rotation affected each model. Values close to 1 indicate little to no effect, while values closer to 0 indicate a strong effect.

A.3 More result

We visualize the rotated image in Figure 9.

A.4 Licenses

Datasets:

- CelebA Xia et al. (2021): Creative Commons BY-NC 4.0 license
- COCO14 Lin et al. (2014): Creative Commons BY-NC 4.0 license
- FFHQ Karras et al. (2019): Creative Commons BY-NC 2.0
- ImageNet Deng et al. (2009): The license status is unclear

Pre-trained models:

- DinoV2 model by Maxime Oquab et al. Oquab et al. (2023): MIT license
- Inception-v3 model by Szegedy et al. Szegedy et al. (2016): Apache V2.0 license

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the Appendix.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: All theoretic information are based on previous works.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the necessary information and the code to reproduce the results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All data and code are accessible to the public.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: They are provided either in the experimental part or the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The convention of this topic normally didn't necessarily report the statistical significance, as the result is quite discriminative.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: They are provided either in the experimental part or the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we are.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: They are stated in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have done our best to incorporate that information.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: Our rotated dataset is a preprocessing of the asset, thus it doesn't bring in new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: No crowdsourcing is involved.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.