Scalable Diffusion Models with State Space Backbone

Zhengcong Fei, Mingyuan Fan, Changqian Yu
Junshi Huang*
{feizhengcong}@gmail.com

Kunlun Inc.

Abstract. This paper presents a new exploration into a category of diffusion models built upon state space architecture. We endeavor to train diffusion models for image data, wherein the traditional U-Net backbone is supplanted by a state space backbone, functioning on raw patches or latent space. Given its notable efficacy in accommodating long-range dependencies, Diffusion State Space Models (DiS) are distinguished by treating all inputs including time, condition, and noisy image patches as tokens. Our assessment of DiS encompasses both unconditional and class-conditional image generation scenarios, revealing that DiS exhibits comparable, if not superior, performance to CNN-based or Transformer-based U-Net architectures of commensurate size. Furthermore, we analyze the scalability of DiS, gauged by the forward pass complexity quantified in Gflops. DiS models with higher Gflops, achieved through augmentation of depth/width or augmentation of input tokens, consistently demonstrate lower FID. In addition to demonstrating commendable scalability characteristics, DiS-H/2 models in latent space achieve performance levels akin to prior diffusion models on class-conditional ImageNet benchmarks at the resolution of 256×256 and 512×512, while significantly reducing the computational burden. The code and models are available at: https://github.com/feizc/DiS.

Keywords: State space · diffusion models · image synthesis

1 Introduction

Diffusion models [4, 26, 60, 64] have emerged as potent deep generative models in recent years [9, 27, 54], for their capacity in high-quality image generation. Their rapid development has led to widespread application across various domains, including text-to-image generation [21, 53–55], image-to-image generation [6, 70, 71], video generation [25, 45, 45], speech synthesis [5, 36], and 3D synthesis [52, 72]. Along with the development of sampling algorithms [10, 32, 41, 42, 62, 66], the revolution of backbones stands as a pivotal aspect in advancement of diffusion models. A representative example is U-Net based on a convolutional neural network (CNN) [26, 63], which has been prominently featured in previous research. The CNN-based UNet is characterized by a group of down-sampling
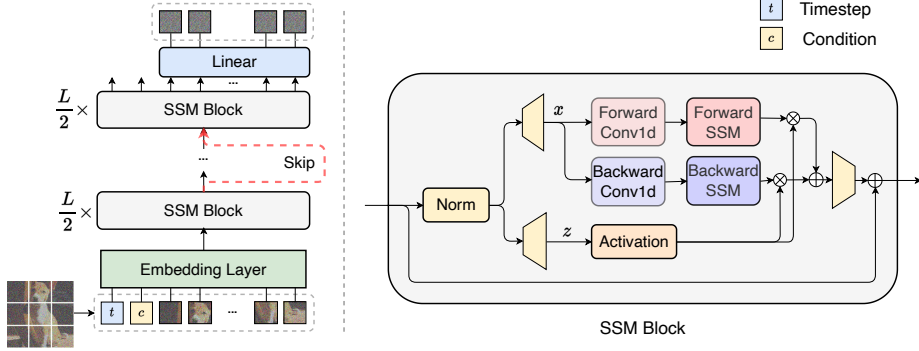
Fig. 1: The proposed state space-based diffusion models. It treats all inputs including the time, condition and noisy image patches as tokens and employs skip connections between shallow and deep layers. Different from original Mamba for text sequence modeling, our SSM block process the hidden states sequence with both forward and backward directions.

blocks, a group of up-sampling blocks, and long skip connections between the two groups [9, 53–55]. Similarly, Transformer-based architecture replaces sampling block with self-attention while keeping the remain untouched [1, 51, 68], resulting in streamlined yet effective performance.

On the other hand, state space models (SSMs) with efficient hardware-aware designs, have shown great potential in the realm of long sequence modeling [19, 20, 22, 29, 49, 59]. As self-attention mechanism in Transformer scales quadratically with the input size, making them resource-intensive when dealing with long-range visual dependencies, *i.e.*, high resolution images. Recent efforts, exampified by the work on Mamba [15, 16], have sought to address by integrating time-varying parameters into the SSM and proposing a hardware-aware algorithm to enable efficient training and inference. The commendable scaling performance of Mamba underscores its promise as a viable avenue for constructing efficient and versatile backbones within the domain of SSMs. Motivated by the successes observed in language modeling with Mamba, a pertinent inquiry arises: whether we can build SSM-based U-Net in diffusion models?

In this paper, we aim to design a simple and general state space-based architecture for diffusion models, referred to as DiS. Following the design principles of [51], DiS treats all inputs including time, condition and noisy image patches as discrete tokens. Crucially, we conduct a systematical ablation study encompassing the incorporation of conditioning factors and the optimization of model architecture across all components. Noteworthy is DiS's adherence to the established best practices of state space models, renowned for their superior scalability in image generation tasks when compared to CNN or Transformers, all while maintaining lower computational overhead.

Experimentally, we assess the performance DiS across both unconditional and class-conditional image generation tasks. In all settings, DiS demonstrate comparative, if not superior, efficay when juxtaposed with CNN-based or Transformer-based U-Nets of a similar size. Additionally, we provide evidence showcasing that DiSs are scalable architectures for diffusion models, where a discernible correlation is observed between network complexity and sample quality. Moreover, experiments yield impressive results, with DiS achieving a comparable FID scores in class-conditional image generation conducted on ImageNet at a resolution of 256×256 and 512×512. This study endeavors to elucidate the criticality of architectural selections within the domain of diffusion models, concurrently providing empirical benchmarks that can inform forthcoming research in generative modeling. Our aspiration is that the insights gleaned from DiS will serve as foundational knowledge for future investigations concerning backbone architectures in diffusion models, ultimately enriching the landscape of generative modeling, particularly in the context of large-scale cross-modality datasets.

2 Methodology

2.1 Preliminaries

Diffusion models. Prior to presenting our architecture, we provide a concise overview of fundamental concepts relevant to our discussion. Diffusion models [26,60] gradually inject noise to data, and subsequently reverse this process to generate data from noise. The noise injection process, also referred to as the forward process, can be formalized as Markov chain as:

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}),$$
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t|\sqrt{\alpha_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

Here $\mathbf{x}_0$ is the data, α_t and β_t represent the noise schedule, ensuring $\alpha_t + \beta_t = 1$. To reverse this process, a Gaussian model $p(\mathbf{x}_{t-1}|x_t) = \mathcal{N}(\mathbf{x}_{t-1}|\mu_t(\mathbf{x}_t), \sigma_t^2 \mathbf{I})$ is employed to approximate the ground truth reverse transition $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$, where the learning is tantamount to a noise prediction task. Formally, a noise prediction network $\epsilon_\theta(\mathbf{x}_t, t)$ is incorporated by minimizing a noise prediction objective, *i.e.*, $\min_\theta \mathbb{E}_{t,\mathbf{x}_0,\epsilon}||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||_2^2$, where t is uniformly distributed between 1 and T. To learn conditional diffusion models, *e.g.*, class-conditional [9] or text-to-image [2,53] models, additional condition information is integrated into the noise prediction objective as:

$$\min_\theta \mathbb{E}_{t,\mathbf{x}_0,\mathbf{c},\epsilon}||\eta - \eta_\theta(\mathbf{x}_t, t, \mathbf{c})||_2^2, \tag{2}$$

where $\mathbf{c}$ is the condition index or its continuous embedding.

State space backbone. State space models are conventionally defined as linear time-invariant systems that builds a map stimulation $x(t) \in \mathbb{R}^N$ to response $y(t) \in \mathbb{R}^N$ by a latent state $h(t) \in \mathbb{R}^N$. The process can be formulated as:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}x(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^N$ denote the state matrix, input matrix, and output matrix, respectively. In the quest to derive the output sequence $y(t)$ at time t, the analytical solution for obtaining $h(t)$ proves to be a formidable challenge. Conversely, real-world data typically manifests in discrete rather than continuous forms. As an alternative, we can discretize the system in Equation 3 as follows:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t, \end{aligned} \tag{4}$$

where $\overline{\mathbf{A}} := \exp(\Delta{\cdot}\mathbf{A})$ and $\overline{\mathbf{B}} := (\Delta{\cdot}\mathbf{A})^{-1}(\exp(\Delta{\cdot}\mathbf{A}) - I){\cdot}\Delta\mathbf{B}$ are the discretized state parameters and Δ is the discretization step size.

While SSMs boast rich theoretical properties, they are often afflicted by high computational costs and numerical instability. These shortcomings have prompted the development of structured state space sequence models (S4), which seek to mitigate these challenges by imposing a structured format on the state matrix $\mathbf{A}$ through the utilization of HIPPO matrices. This structural enhancement has led to notable advancements in both performance and efficiency. Notably, S4 has exhibited a substantial performance advantage over Transformers, which necessitates effective modeling of long-range dependencies [19]. More recently, Mamba [16] further advance its potential through an input-dependent selection mechanism and a faster hardware-aware algorithm.

2.2   Model Structure Design

We introduce Diffusion State Space Models, denoted as DiS, a simple and general architecture for diffusion models in image generation. Specifically, DiS parameterizes the noise prediction network $\epsilon_\theta(\mathbf{x}_t, t, \mathbf{c})$, which takes the timestep t, condition $\mathbf{c}$ and noised image $\mathbf{x}_t$ as inputs and predicts the noise injected into $\mathbf{x}_t$. Our objective is to closely adhere to the advanced state space architecture to preserve its scalability characteristics, hence, DiS is based on the bidirectional Mamba [16,73] architecture which operates on sequences of tokens. Figure 1 provides an overview of the complete DiS architecture. In this section, we expound upon the forward pass of DiS, as well as the components of the design space of the DiS class.

Image patchnify. The first layer of DiS undertakes transformation of input image $\mathbf{I} \in \mathbb{R}^{H \times W \times C}$ into flatten 2-D patches $\mathbf{X} \in \mathbb{R}^{J \times (p^2 \cdot C)}$. Subsequently,

converts it into a sequence of J tokens, each of dimension D, by linearly embedding each patch in the input. Consistent with [11], we apply learnable positional embeddings to all input tokens. The number of tokens J created by patchify is determined by the hyperparameter patch size p to $\frac{H \times W}{p^2}$. The patchnify supports both raw pixel and latent space. We set $p = 2, 4, 8$ to the design space.

SSM block. After embedding layer, the input tokens are processed by a sequence of SSM blocks. In addition to noised image inputs, diffusion models sometimes process additional conditional information such as noise timesteps t, condition **c** such as class labels or natural language. Given that original Mamba block is designed for 1-D sequence, we resort to [73], which incorporates bidirectional sequence modeling tailed for vision tasks. The designs introduce subtle yet pivotal modifications to the standard SSM block design. As shown in Figure 1 right part, forward of SSM blocks combine both forward and backward directions.

Skip connection. Given a series of L SSM blocks, we categorize the stack SSM block into first half $\lfloor \frac{L}{2} \rfloor$ shallow group, one middle layer, and second half $\lfloor \frac{L}{2} \rfloor$ deep group. Let $\mathbf{h}_{shallow}, \mathbf{h}_{deep} \in \mathbb{R}^{J \times D}$ be the hidden states from the main branch and long skip branch respectively. We consider directly concatenating them and perform a linear projection, *i.e.*, `Linear(Concat(`$\mathbf{h}_{shallow}, \mathbf{h}_{deep}$`))`, before feeding them to the next SSM block.

Linear decoder. After the final SSM block, we need to decode our sequence of hidden states into an output noise prediction and diagonal covariance prediction. Both of these outputs retain a shape identical to the original spatial input. We use a standard linear decoder, *i.e.*, we apply the final layer norm and linearly decode each token into a $p^2 \cdot C$ tensor. Finally, we rearrange the decoded tokens into their original spatial layout to get the predicted noise and covariance.

Condition incorporation. To effectively integrate additional conditions, we adopt a straightforward strategy of appending the vector embeddings of timestep t and condition **c** as two supplementary tokens in the input sequence. These tokens are treated equivalently to the image tokens, akin to the approach with cls tokens in Vision Transformers [11]. This approach enables the utilization of SSM blocks without necessitating any modifications. Following the final block, the conditioning tokens are removed from the sequence. We also explore to adaptive normalization layer, where replace standard norm layer with adaptive norm layer. That is, rather than directly learn dimension-wise scale and shift parameters, we regress them from the sum of the embedding vectors of t and **c**, which will be discussed in the Experiments Section.

Table 1: Scaling law model size. The model sizes and detailed hyperparameters settings for scaling experiments.

	#Params	#Blocks L	Hidden dim. D	#Expand E	Gflops
Small	28.4M	25	384	2	0.43
Base	119.1M	25	768	2	1.86
Medium	229.4M	49	768	2	3.70
Large	404.0M	49	1024	2	6.57
Huge	900.6M	49	1536	2	14.79

2.3   Computation Analysis

In summary, the hyper-parameters of our architecture encompass the following: the number of blocks L, hidden state dimension D, expanded state dimension E, and SSM dimension N. Various configurations of DiS are delineated in Table 1. They cover a wide range of model sizes and flop allocations, from 28M to 900M, thus affording comprehensive insights into the scalability performance. Aligned with [51], Gflop metric is evaluated in 32×32 unconditional image generation with patch size $p = 4$ with `thop` python package. Following [16] we also set the SSM dimension of all model variants to 16.

Both SSM block in DiS and self-attention in Transformer play a key role in modeling long context adaptively. We further provide a theoretical analysis pertaining to computation efficiency. Given a sequence $\mathbf{X} \in \mathbb{R}^{1 \times J \times D}$ and the default setting $E = 2$, the computation complexity of a self-attention and SSM operation are delineated as:

$$\mathcal{O}(\text{SA}) = 4JD^2 + 2J^2D, \tag{5}$$

$$\mathcal{O}(\text{SSM}) = 3J(2D)N + J(2D)N^2, \tag{6}$$

where we can see that self-attention is quadratic to sequence length J, and SSM is linear with respect to sequence length J. It is noteworthy that N is a fixed parameter, typically set to 16 by default. This computational efficiency renders DiS amenable to scalability in scenarios necessitating generation with large sequence lengths, such as gigapixel applications.

3   Experiments

We delve into the design space and scrutinize the scaling properties of our DiS model class. Our model is named according to their configs and patch size p; for instance, DiS-L/2 refers to the Large version config and $p = 2$.

3.1   Experimental Settings

Datasets. For unconditional image learning, we consider CIFAR10 [37], which comprises 50K training images, and CelebA 64x64 [40], which contains 162,770
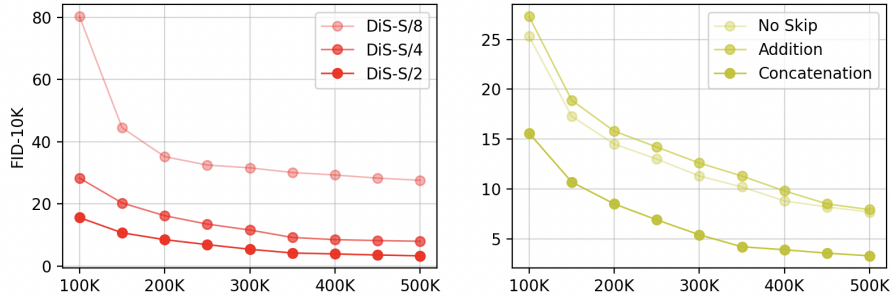
Fig. 2: Ablation experiments with DiS-S/2 on CIFAR10 dataset. We report FID metrics on 10K generated samples. (a) **Patch size**. A smaller patch size can improve the generation performance. (b) **Long skip**. Combining the long skip branch can accelerate the training as well as optimize generated results.

training images of human faces. For class-conditional image learning, we consider ImageNet [8] at 256×256 and 512×512 resolutions, which contains 1,281,167 training images across 1,000 different classes. The only data augmentation is horizontal flips. We train 500K iterations on CIFAR10 and CelebA 64×64 with a batch size of 128. We train 500K and 1M iterations on ImageNet 256×256 and 512×512, with a batch size of 1024.

Implementation details.   We use the AdamW optimizer [33] without weight decay across all datasets. We maintain a learning rate of 1e-4 with a cosine schedule. In our early experiments, we tried learning rates ranging from 1e-4 to 5e-4, and found that loss may become NAN if the learning rate set is larger than 3e-4. Consistent with practice in literature [51], we utilize an exponential moving average of DiS weights over training with a decay of 0.9999. All results were reported using the EMA model. Our models are trained on Nvidia A100 GPU. When trained on ImageNet dataset at resolution of 256×256 and 512×512, we adopt classifier-free guidance [28] following [54] and use a off-the-shelf pre-trained variational autoencoder (VAE) model [35] from Stable Diffusion [54] provided in huggingface[1]. The VAE encoder has a downsampling factor of 8. We retrain diffusion hyperparameters from [51], using a $t_{max} = 1000$ linear variance schedule ranging from 1×10^{-4} to 2×10^{-2} and parameterization of the covariance.

Evaluation metrics. We measure image generation performance with Frechet Inception Distance (FID) [24], a widely adopted metric for assessing the quality of generated images. We follow convention when comparing against prior works and report FID-50K using 250 DDPM sampling steps [50] following the process
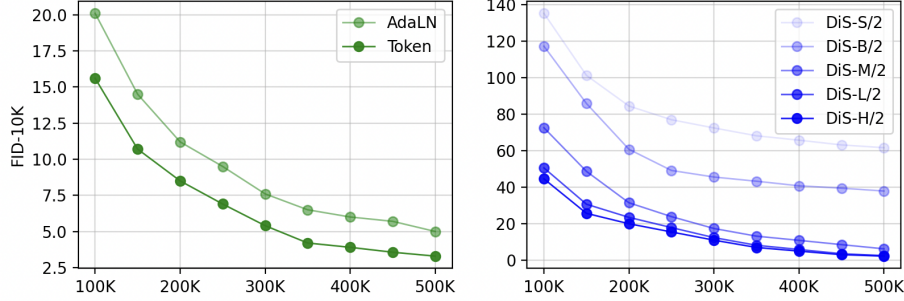
---

[1] https://huggingface.co/stabilityai/sd-vae-ft-ema

Fig. 3: Model analysis for different designs. (a) **Variants of condition incorporation**. Feed time as token into the networks is effective. (b) **Model parameters scaling**. As we expected, holding patch size constant, increasing the model size can consistently improve the generation performance.

of [9]. Notably, the FID values presented in this section are calculated without classifier-free guidance, unless specified otherwise. We additionally report Inception Score [56], sFID [47] and Precision/Recall [38] as secondary metrics.

3.2   Model Analysis

In this section, we perform a systematical empirical investigation into the fundamental components of DiS models. Specifically, we ablate on the CIFAR10 dataset, evaluate the FID score every 50K training iterations on 10K generated samples, instead of 50K samples for efficiency identical to [1], and determine the optimal default implementation details.

Effect of patch size. We train patch size range over (8, 4, 2) with DiS-S configuration on the CIFAR10 dataset. As depicted in Figure 2(a) shows FID metric fluctuates in response to patch size decrease when maintaining model size consistent. We observe that discernible FID enhancements are observed throughout the training process by augmenting the number of tokens processed by DiS, holding parameters approximately fixed. Consequently, optimal performance necessitates a smaller patch size, such as 2. We postulate that this requirement stems from the inherently low-level nature of the noise prediction task in diffusion models, which favors smaller patches, in contrast to higher-level tasks like classification. Moreover, given the computational costs associated with using small patch sizes for high-resolution images, an alternative approach involves transforming these images into low-dimensional latent representations, subsequently also modeled using the DiS series.

Effect of long skip. For evaluating the efficiency of skipping operation, we examine three variants including: (**i**) Concatenation, *i.e.*, Linear(Concat($\mathbf{h}_{shallow}$,

Table 2: Benchmarking unconditional image generation on CIFAR10. DiS-S/2 model obtains comparable results with fewer parameters.

Table 3: Benchmarking unconditional image generation on CelebA 64×64. DiS-S/2 maintains a superior generation performance in small model settings.

Unconditional CIFAR10		
Model	#Params	FID↓
Diff. based on U-Net		
DDPM [26]	36M	3.17
EDM [30]	56M	1.97
Diff. based on Transformer		
GenViT [68]	11M	20.20
U-ViT-S/2 [1]	44M	3.11
Diff. based on SSM		
DiS-S/2	28M	3.25

Unconditional CelebA 64×64		
Model	#Params	FID↓
Diff. based on U-Net		
DDIM [61]	79M	3.26
Soft Trunc. [31]	62M	1.90
Diff. based on Transformer		
U-ViT-S/4 [1]	44M	2.87
Diff. based on SSM		
DiS-S/2	28M	2.05

$\mathbf{h}_{deep}$)); (ii) Addition, *i.e.*, $\mathbf{h}_{shallow} + \mathbf{h}_{deep}$; (iii) No skip connection. As show in Figure 2(b), directly adding the hidden states from the shallow and deep layers does not yield any discernible benefits. Given that the SSM block already incorporates residual skip connections internally, the hidden states of deep layer inherently encapsulates some information from the shallow layer in a linear fashion. Consequently, the addition operation in the long skip merely amplifies the contribution of the shallow layer's hidden state, without fundamentally altering the network's behavior. In contrast, employing concatenation involves a learnable linear projection on the shallow hidden states and effectively enhances performance compared to the absence of a long skip connection.

Effect of condition combination. We explore two ways to integrate the conditional timestep t into the network: (**i**) treat timestep t as a token and directly concatenate it with image patches in prefix. (**ii**) incorporate the time embedding after the layer normalization in the SSM block, which is similar to the adaptive group normalization [9] used in U-Net. The second way is referred to as adaptive layer normalization (AdaLN). Formally, $\text{AdaLN}(h, s) = y_s \text{LayerNorm}(h) + y_b$, where h is a hidden state inside an SSM block, and y_s and y_b are obtained from a linear projection of the time embedding. As illustrated in Figure 3 (a), while simple and straightforward, the first way that treats time as a token performs better than AdaLN.

Scaling model size. We investigate scaling properties of DiS by studying the effect of depth, *i.e.*, number of SSM layers, width, e.g. the hidden size. Specifically, we train 5 DiT models on the ImageNet dataset with a resolution of 256×256, spanning model configurations from small to huge as detailed in Table 1, denoted as (S, B, M, L, H) for simple. As shown in Figure 3 (b), the performance

Table 4: Benchmarking class-conditional image generation on ImageNet 256×256. DiS-H/2 achieves state-of-the-art FID metrics towards best competitors.

Class-Conditional ImageNet 256×256					
Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑
GAN					
BigGAN-deep [3]	6.95	7.36	171.4	0.87	0.28
StyleGAN-XL [57]	2.30	4.02	265.12	0.78	0.53
Diff. based on U-Net					
ADM [9]	10.94	6.02	100.98	0.69	0.63
ADM-U	7.49	5.13	127.49	0.72	0.63
ADM-G	4.59	5.25	186.70	0.82	0.52
ADM-G, ADM-U	3.94	6.14	215.84	0.83	0.53
CDM [27]	4.88	-	158.71	-	-
LDM-8 [54]	15.51	-	79.03	0.65	0.63
LDM-8-G	7.76	-	209.52	0.84	0.35
LDM-4	10.56	-	103.49	0.71	0.62
LDM-4-G	3.60	-	247.67	0.87	0.48
VDM++ [34]	2.12	-	267.70	-	-
Diff. based on Transformer					
U-ViT-H/2 [1]	2.29	5.68	263.88	0.82	0.57
DiT-XL/2 [51]	2.27	4.60	278.24	0.83	0.57
Diff. based on SSM					
DiS-H/2	2.10	4.55	271.32	0.82	0.58

improves as the depth increase from 25 to 49. Similarly, increasing the width from 384 to 768 yields performance gains. Notably, the model's performance has not yet reached convergence. Overall, across all five configurations, significant enhancements in FID are observed across all training stages by augmenting the depth and width of the SSM structure.

3.3   Main Results

Unconditional image generation. We conduct a comparative analysis between the DiT-S/2 model and prior diffusion models based on U-Net as well as Transformer-based diffusion models. Consistent with previous literature [1], we evaluate image quality using the FID score, based on 50K generated samples. As summarized in Tables 2 and 3, DiS exhibits comparable performance to U-Net and U-ViT-S/2 on unconditional CIFAR10 and CelebA 64×64 datasets, meanwhile significantly outperforming GenViT. Moreover, notably, our model achieves approximately a 50% reduction in model parameters.

Class-conditional image generation. To demonstrate performance in latent space, where images are initially converted to their latent representations before applying diffusion models, we conducted additional experiments. The evaluation

Table 5: Benchmarking class-conditional image generation on ImageNet 512×512. DiS-H/2 demonstrates a promising performance compared with both CNN-based and Transformer-based UNet for diffusion.

Class-Conditional ImageNet 512×512					
Model	FID↓	sFID↓	IS↑	Precision↑	Recall↑
GAN					
BigGAN-deep [3]	8.43	8.13	177.90	0.88	0.29
StyleGAN-XL [57]	2.41	4.06	267.75	0.77	0.52
Diff. based on U-Net					
ADM [9]	23.24	10.19	58.06	0.73	0.60
ADM-U	9.96	5.62	121.78	0.75	0.64
ADM-G	7.72	6.57	172.71	0.87	0.42
ADM-G, ADM-U	3.85	5.86	221.72	0.84	0.53
VDM++ [34]	2.65	-	278.10	-	-
Diff. based on Transformer					
U-ViT-H/4 [1]	4.05	6.44	263.79	0.84	0.48
DiT-XL/2 [51]	3.04	5.02	240.82	0.84	0.54
Diff. based on SSM					
DiS-H/2	2.88	4.74	272.33	0.84	0.56

results are listed in Tables 4 and 5. On the class-conditional ImageNet 256×256 dataset, our DiS-H/2 achieves an FID of 2.10, surpassing all prior models. Notably, DiS-H/2 also outperforms DiT-XL/2, a competitive diffusion model employing a transformer as its backbone. Furthermore, on the class-conditional ImageNet 512×512 dataset, our DiS-H/2 surpasses ADM-G, a model directly modeling image pixels. Finally, we also observe that DiS-H/2 achieves higher recall values across all tested classifier-free guidance scales compared to LDM-4 and LDM-8 in various settings.

3.4   Case Study

Figure 4 showcases a selection of samples from ImageNet datasets at resolutions of 256×256 and 512×512, along with random samples from other datasets, demonstrating clear semantics and high-quality generation. For further exploration, additional generated samples, including both class-conditional and random ones, are available on the project page.

4   Related Works

4.1   State Space Backbone for Sequence Modeling

State space models, a recent addition to the realm of deep learning, have garnered attention for their capacity to transform state spaces [19, 20, 31]. Drawing

Fig. 4: Image results generated from DiS model. Selected samples on ImageNet 256×256, and random samples on CIFAR10, CelebA 64×64. We can see that DiS can generate high-quality images while keeping integrated condition alignment.

inspiration from continuous state space models in control systems and leveraging the HiPPO initialization [17], LSSL [20] has demonstrated promise in addressing long-range dependency issues. However, the computational and memory demands associated with the state representation render LSSL impractical for real-world applications. To mitigate this challenge, S4 [19] proposes normalizing parameters into a diagonal structure. Subsequently, various iterations of structured state-space models have emerged, featuring diverse structures such as complex-diagonal configurations [18, 22], multiple-input multiple-output supporting [59], decomposition of diagonal plus low-rank operations [23], selection mechanism [16]. These models have been integrated into large representation models [15, 43, 44], primarily focusing on applications involving long-range and causal data such as language and speech, including tasks like language under-

standing [43] and content-based reasoning [16]. Seminal work [67] is most related to us, they analyze attention-free structure for image synthesis. In this paper, we focus on the scaling properties of advanced Mamba structure when used as the backbone of diffusion models of images.

4.2 Image Generation in Diffusions

Diffusion models [7, 9, 26, 48, 60] represent a category of generative probabilistic models designed to approximate data distributions and facilitate straightforward sampling procedures. Typically, these models operate by taking a Gaussian noise input and iteratively denoising it through a series of gradual steps until it conforms to the target distribution. The specifics of the denoising process, including the number of steps and the transformation parameterization, vary across different studies [39, 41, 42, 60, 61]. Recently, diffusion models have emerged as leading generators renowned for their ability to learn intricate distributions and produce diverse, high-quality samples. These models have found successful application across various domains, including images [9, 12–14, 48, 53–55], video [58], 3D scenes [46], motion sequences [65, 69], and so on. We follow the path of image generation with diffusion, but consider the impact of high-efficient structures.

5 Conclusion

This paper presents Diffusion State Space models (DiS), a simple and general state space-based framework for image generation using diffusion models. DiS adopts a unified approach to handle all inputs, including time, conditions, and noisy image patches, treating them as concatenated tokens. Experimental results indicate that DiS compares favorably, if not surpasses, prior CNN-based or Transformer-based U-Net models while inheriting the remarkable scalability characteristic of the state space model class. We posit that DiS can offer valuable insights for future investigations into backbones within diffusion models and contribute to advancements in generative modeling across large-scale multimodal datasets. Given the encouraging scalability outcomes presented in this study, future endeavors should focus on further scaling DiS to larger models and token counts.

References

1. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A vit backbone for diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22669–22679 (2023) 2, 8, 9, 10, 11
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023) 3
3. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018) 10, 11

4. Cao, H., Tan, C., Gao, Z., Xu, Y., Chen, G., Heng, P.A., Li, S.Z.: A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering (2024) 1

5. Chen, N., Zhang, Y., Zen, H., Weiss, R.J., Norouzi, M., Chan, W.: Wavegrad: Estimating gradients for waveform generation. arXiv preprint arXiv:2009.00713 (2020) 1

6. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021) 1

7. Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M.: Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023) 13

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009) 7

9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021) 1, 2, 3, 8, 9, 10, 11, 13

10. Dockhorn, T., Vahdat, A., Kreis, K.: Score-based generative modeling with critically-damped langevin diffusion. arXiv preprint arXiv:2112.07068 (2021) 1

11. 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. arXiv preprint arXiv:2010.11929 (2020) 5

12. Duan, X., Cui, S., Kang, G., Zhang, B., Fei, Z., Fan, M., Huang, J.: Tuning-free inversion-enhanced control for consistent image editing. arXiv preprint arXiv:2312.14611 (2023) 13

13. Fei, Z., Fan, M., Huang, J.: Gradient-free textual inversion. arXiv preprint arXiv:2304.05818 (2023) 13

14. Fei, Z., Fan, M., Huang, J., Wei, X., Wei, X.: Progressive denoising model for fine-grained text-to-image generation. arXiv preprint arXiv:2210.02291 (2022) 13

15. 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 preprint arXiv:2212.14052 (2022) 2, 12

16. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023) 2, 4, 6, 12, 13

17. Gu, A., Dao, T., Ermon, S., Rudra, A., Ré, C.: Hippo: Recurrent memory with optimal polynomial projections. Advances in neural information processing systems **33**, 1474–1487 (2020) 12

18. Gu, A., Goel, K., Gupta, A., Ré, C.: On the parameterization and initialization of diagonal state space models. Advances in Neural Information Processing Systems **35**, 35971–35983 (2022) 12

19. Gu, A., Goel, K., Ré, C.: Efficiently modeling long sequences with structured state spaces. arXiv preprint arXiv:2111.00396 (2021) 2, 4, 11, 12

20. 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. Advances in neural information processing systems **34**, 572–585 (2021) 2, 11, 12

21. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10696–10706 (2022) 1

22. Gupta, A., Gu, A., Berant, J.: Diagonal state spaces are as effective as structured state spaces. Advances in Neural Information Processing Systems **35**, 22982–22994 (2022) 2, 12

23. Hasani, R., Lechner, M., Wang, T.H., Chahine, M., Amini, A., Rus, D.: Liquid structural state-space models. arXiv preprint arXiv:2209.12951 (2022) 12

24. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017) 7

25. Ho, J., Chan, W., Saharia, C., Whang, J., Gao, R., Gritsenko, A., Kingma, D.P., Poole, B., Norouzi, M., Fleet, D.J., et al.: Imagen video: High definition video generation with diffusion models. arXiv preprint arXiv:2210.02303 (2022) 1

26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 1, 3, 9, 13

27. Ho, J., Saharia, C., Chan, W., Fleet, D.J., Norouzi, M., Salimans, T.: Cascaded diffusion models for high fidelity image generation. The Journal of Machine Learning Research **23**(1), 2249–2281 (2022) 1, 10

28. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022) 7

29. Kalman, R.E.: A new approach to linear filtering and prediction problems (1960) 2

30. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. Advances in Neural Information Processing Systems **35**, 26565–26577 (2022) 9

31. Kim, D., Shin, S., Song, K., Kang, W., Moon, I.C.: Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. arXiv preprint arXiv:2106.05527 (2021) 9, 11

32. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Advances in neural information processing systems **34**, 21696–21707 (2021) 1

33. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 7

34. Kingma, D.P., Gao, R.: Understanding diffusion objectives as the elbo with simple data augmentation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023) 10, 11

35. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 7

36. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. arXiv preprint arXiv:2009.09761 (2020) 1

37. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009) 6

38. Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., Aila, T.: Improved precision and recall metric for assessing generative models. Advances in Neural Information Processing Systems **32** (2019) 8

39. Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022) 13

40. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015) 6

41. Lu, C., Zheng, K., Bao, F., Chen, J., Li, C., Zhu, J.: Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In: International Conference on Machine Learning. pp. 14429–14460. PMLR (2022) 1, 13

42. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in Neural Information Processing Systems **35**, 5775–5787 (2022) 1, 13
43. Ma, X., Zhou, C., Kong, X., He, J., Gui, L., Neubig, G., May, J., Zettlemoyer, L.: Mega: moving average equipped gated attention. arXiv preprint arXiv:2209.10655 (2022) 12, 13
44. Mehta, H., Gupta, A., Cutkosky, A., Neyshabur, B.: Long range language modeling via gated state spaces. arXiv preprint arXiv:2206.13947 (2022) 12
45. Mei, K., Patel, V.: Vidm: Video implicit diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9117–9125 (2023) 1
46. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4328–4338 (2023) 13
47. Nash, C., Menick, J., Dieleman, S., Battaglia, P.W.: Generating images with sparse representations. arXiv preprint arXiv:2103.03841 (2021) 8
48. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning. pp. 8162–8171. PMLR (2021) 13
49. Orvieto, A., Smith, S.L., Gu, A., Fernando, A., Gulcehre, C., Pascanu, R., De, S.: Resurrecting recurrent neural networks for long sequences. arXiv preprint arXiv:2303.06349 (2023) 2
50. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in gan evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022) 7
51. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4195–4205 (2023) 2, 6, 7, 10, 11
52. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 1
53. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 1, 2, 3, 13
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 1, 2, 7, 10, 13
55. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (2022) 1, 2, 13
56. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29** (2016) 8
57. Sauer, A., Schwarz, K., Geiger, A.: Stylegan-xl: Scaling stylegan to large diverse datasets. In: ACM SIGGRAPH 2022 conference proceedings. pp. 1–10 (2022) 10, 11
58. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022) 13
59. Smith, J.T., Warrington, A., Linderman, S.W.: Simplified state space layers for sequence modeling. arXiv preprint arXiv:2208.04933 (2022) 2, 12

60. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. PMLR (2015) 1, 3, 13

61. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 9, 13

62. Song, Y., Durkan, C., Murray, I., Ermon, S.: Maximum likelihood training of score-based diffusion models. Advances in Neural Information Processing Systems **34**, 1415–1428 (2021) 1

63. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems **32** (2019) 1

64. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020) 1

65. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022) 13

66. Vahdat, A., Kreis, K., Kautz, J.: Score-based generative modeling in latent space. Advances in Neural Information Processing Systems **34**, 11287–11302 (2021) 1

67. Yan, J.N., Gu, J., Rush, A.M.: Diffusion models without attention. arXiv preprint arXiv:2311.18257 (2023) 13

68. Yang, X., Shih, S.M., Fu, Y., Zhao, X., Ji, S.: Your vit is secretly a hybrid discriminative-generative diffusion model. arXiv preprint arXiv:2208.07791 (2022) 2, 9

69. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16010–16021 (2023) 13

70. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023) 1

71. Zhao, M., Bao, F., Li, C., Zhu, J.: Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. Advances in Neural Information Processing Systems **35**, 3609–3623 (2022) 1

72. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12588–12597 (2023) 1

73. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. arXiv preprint arXiv:2401.09417 (2024) 4, 5