SSM MEETS VIDEO DIFFUSION MODELS: EFFICIENT VIDEO GENERATION WITH STRUCTURED STATE SPACES

Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki & Yutaka Matsuo
The University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, Japan.
{yuta.oshima, taniguchi, masa, matsuo}@weblab.t.u-tokyo.ac.jp

ABSTRACT

Given the remarkable achievements in image generation through diffusion models, the research community has shown increasing interest in extending these models to video generation. Recent diffusion models for video generation have predominantly utilized attention layers to extract temporal features. However, attention layers are limited by their memory consumption, which increases quadratically with the length of the sequence. This limitation presents significant challenges when attempting to generate longer video sequences using diffusion models. To overcome this challenge, we propose leveraging state-space models (SSMs). SSMs have recently gained attention as viable alternatives due to their linear memory consumption relative to sequence length. In the experiments, we first evaluate our SSM-based model with UCF101, a standard benchmark of video generation. In addition, to investigate the potential of SSMs for longer video generation, we perform an experiment using the MineRL Navigate dataset, varying the number of frames to 64, 200, and 400. In these settings, our SSM-based model can considerably save memory consumption for longer sequences, while maintaining competitive FVD scores to the attention-based models.

1 INTRODUCTION

Research on video generation diffusion models (Sohl-Dickstein et al., 2015; Nichol & Dhariwal, 2021; Ho et al., 2020) is cutting-edge in the field of deep generative models. The success of image generation using diffusion models, notably Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020), has sparked a surge in research on applying diffusion models to video generation. This trend has been exemplified by the emergence of video diffusion models (VDMs) (Ho et al., 2022b). By harnessing the substantial representational capacity inherent in diffusion models, their application to video generation has showcased impressive performance in modeling the dynamic and intricate nature of video content (Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a).

However, research on diffusion-model-based video generation faces significant challenges in terms of computational complexity with respect to temporal sequence length. In diffusion-model-based approaches to video generation, attention mechanisms (Vaswani et al., 2017) are employed to capture temporal relationships (Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; Blattmann et al., 2023). In early studies on diffusion models for video generation, such as VDMs, to capture temporal relationships across video frames, temporal attention layers were added subsequent to spatial attention layers within the architecture of diffusion models for image generation, as described in Figure 1 (a). However, the memory demands of attention layers, which scale with the square of the sequence length, present substantial challenges for extending these models to handle longer sequences.

Recently, state-space models (SSMs) (Gu et al., 2020; 2021; 2022; Smith et al., 2023) have been identified as promising alternatives to attention mechanisms. A pioneering work by Gu et al. (2021) has enabled SSMs to capture long-term dependencies of sequential data, and their model, named the structured state space sequence model (S4), demonstrates superior performance in various benchmarks of sequence modeling. In contrast to attention mechanisms, SSMs can handle sequential data with

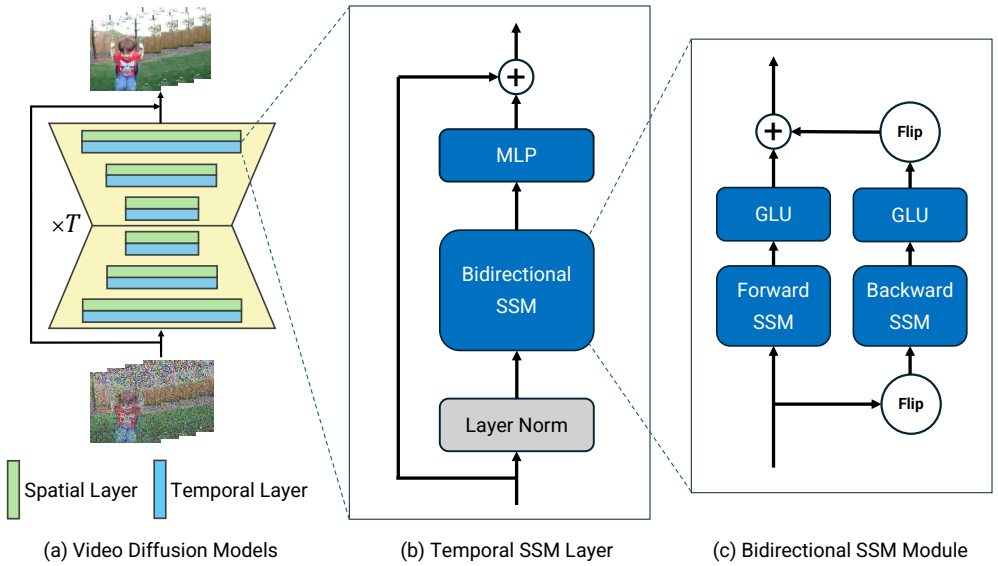


Figure 1: (a) U-Net based video diffusion models consist of spatial layers and temporal layers. (b) In our temporal SSM layer, we replace an attention module with a bidirectional SSM module + a dual-layer MLP in a traditional temporal layer in VDMs. (c) Details of a bidirectional SSM in our temporal SSM layer. Following common practices, GLU is used as an activation function. Element-wise summation is utilized to aggregate features from bidirectional SSMs.

linear complexities, so they are expected to overcome the fundamental limitation of attention-based models in many sequence modeling tasks.

However, in the field of video generation, the application of SSMs has not been explored. This is because a methodology to effectively incorporate SSMs into video generation models has not been established. In fact, we have empirically observed that a naive approach that directly replace temporal attention layers of VDMs with SSMs works much worse than the original attention-based VDMs. To bridge this gap, in this paper, we investigate an effective approach to incorporate SSMs with video diffusion models. Our proposed model is summarized in Figure 1. A key insight of our investigation is that using bidirectional SSMs is essential to achieve good performance for SSM-based VDMs. In addition, to capture complex nonlinear dynamics of video sequences, adding a multi-layer perceptron (MLP) after the bidirectional SSM is also very effective to improve the generative performance.

In the experiments, we demonstrate that by substituting temporal attention in video diffusion models with our temporal SSM layers, it is feasible to show competitive or even better generative performance in benchmarks of video generation, e.g., UCF101 (Soomro et al., 2012), in terms of the Fréchet Video Distance (FVD) (Unterthiner et al., 2018). Moreover, we additionally observe that our SSM-based model can be successfully trained with 400-frame videos of MineRL Navigate (Guss et al., 2019; Saxena et al., 2021) with eight A100 GPUs, while attention-based VDM cannot due to huge memory consumption.

2 Diffusion Models for Video Generation

In diffusion models, the forward process involves progressively diminishing the original data signal, $\mathbf{x}_0$, by gradually introducing Gaussian noise as the diffusion time, t, advances. The generation process in diffusion models is the reverse process. This process starts with pure Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ and gradually reconstructing the data towards the original $\mathbf{x}_0$. During the reverse process, each step $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ is modeled using a neural network. In terms of the architecture of diffusion models, 2D U-Net (Ronneberger et al., 2015) architectures are commonly used for image data. In 2D U-Net-based models, spatial attention layers are incorporated between the convolutional layers. These spatial attention layers enhance the ability to focus on relevant spatial features and improve the quality of the generated images.

To generate videos, diffusion models need to encapsulate both spatial and temporal features across frames. While DDPMs typically comprise a combination of U-Net and spatial attention layer, their capability is predominantly confined to spatial feature capture. To address this limitation, Video Diffusion Models (VDMs) (Ho et al., 2022b) were introduced as an initial attempt into video generation using diffusion models. By incorporating mechanisms to capture temporal dynamics within DDPMs, VDMs enhance their capability to capture temporal features (Figure 1(a)). Temporal attention layer is commonly used in video generation diffusion models, such as VDMs, to leverage time-series dependencies. However, temporal attention requires memory proportional to the square of the sequence length, which imposes limitations on the maximum length of video sequences that can be generated at once. In our study, we adopt VDMs as a baseline to explore the existing challenges and potential improvements in video generation diffusion models.

3 METHOD

In this section, we propose the architecture of a temporal SSM (state-space model) layer for use in diffusion models for videos. Recent diffusion model-based video generation techniques capture temporal features through temporal attention layers, incurring memory costs proportional to the square of the sequence length. Recently, SSMs have emerged as a promising alternative to attention, offering linear memory costs with respect to time (Gu et al., 2020; 2021; 2022; Smith et al., 2023). We first review the recent advancements in SSMs in prior works, followed by a detailed description of our proposed temporal SSM layer architecture for video generation diffusion models.

3.1 STATE SPACE MODELS

Unlike the temporal attention commonly used in video diffusion models, state-space models (SSMs) enable the processing of time series with spatial complexities proportional to the sequence length. Recent studies proposed SSMs that could process inputs in parallel unlike recurrent neural networks (RNNs) (Chung et al., 2014). SSMs are widely used as sequence models that define a mapping from one-dimensional input signals $u(t)$ to one-dimensional output signals $y(t)$:

$$\mathbf{s}_k = \bar{\mathbf{A}}\mathbf{s}_{k-1} + \bar{\mathbf{B}}u_k, y_k = \bar{\mathbf{C}}\mathbf{s}_k. \tag{1}$$

In this study, we consider discrete-time SSMs. Unlike RNNs but akin to attention mechanisms, linear SSMs are capable of parallel computations through discrete convolutions, facilitated by the Fast Fourier Transform. The memory requirements and parallelization capability of the modules managing the time series data are concisely summarized in Table 1. In our research, we chose S4D Gu et al. (2022) as our backbone model for our temporal SSM layer.

As delineated in Equations 1, SSMs are inherently designed to independently handle single-input, single-output systems. Therefore, when employing SSMs to manage multidimensional inputs, it is common practice to append a structure to capture dependencies (such as GLU (Dauphin et al., 2017)) between different dimensions of the output after the SSM (Gu et al., 2021; 2022) (see Figure 1 (c)).

3.2 TEMPORAL SSM LAYER FOR DIFFUSION MODEL-BASED VIDEO GENERATION

We incorporate state-space models (SSMs) within the temporal layers for the video generation diffusion model. The structure of our proposed temporal SSM layer (Figure 1 (b)), is a key element of our study and warrants detailed discussion. This design takes cues from the structure of the temporal attention layer in VDMs, which consists of Layer Normalization (Ba et al., 2016) followed by an attention mechanism and a skip connection (He et al., 2016).

In our model, we replace the self-attention component with an SSM. We adopt a bidirectional structure, drawing from practices in Graves & Schmidhuber (2005); Wang et al. (2022); Yan et al. (2023). This choice is motivated by the inherent limitation of a single SSM, which is typically restricted to capturing unidirectional temporal transitions. By adopting a bidirectional approach, SSMs can more comprehensively understand the temporal dynamics in video data, addressing the constraints of traditional unidirectional SSMs. We employ element-wise summation for the integration of forward and backward SSM outputs, rather than employing the methods taken by Wang et al. (2022); Yan et al. (2023). Experimentally, this methodological choice is informed by its demonstrable

Module Type	Parallel Computation	Memory Usage
RNN	×	BLD
Attention	√	$B(L^2 + DL)$
Linear Attention	√	$B(D^2 + DL)$
SSM	√	BLD

Table 1: Comparison of time-series handling modules in terms of parallelization capability and memory usage related to batch size (B), sequence length (L) and hidden dimension (D).

superiority in boosting the generative capabilities of our temporal layer for video diffusion model to the concatenation-based approach used in Yan et al. (2023).

Additionally, we recognize that SSM, while effective, has limitations in integrating information across different dimensions, demonstrated in Equation 1. To overcome this, we supplement the SSM with a multi-layer perceptron (MLP) after the bidirectional SSM module (Figure 1(b)). Our experiments demonstrate that this addition significantly enhances the model's performance, proving its importance in the overall architecture.

Given an input $\mathbf{X} \in \mathbb{R}^{(B\times H\times W)\times L\times C}$, where L is the sequence length, C is channel size, H is height, W is width of the input image, our proposed temporal SSM layer operates as follows:

$$
\begin{aligned}
\mathbf{H} &= \text{LayerNorm}(\mathbf{X}),\\
\mathbf{F} &= \text{GLU}(\text{SSM}_{\text{forward}}(\mathbf{H})),\\
\mathbf{B} &= \text{GLU}(\text{SSM}_{\text{backward}}(\text{Flip}(\mathbf{H}))),\\
\mathbf{U} &= \mathbf{F} + \text{Flip}(\mathbf{B}),\\
\mathbf{O} &= \text{MLP}(\mathbf{U}) + \mathbf{H}.
\end{aligned}
$$

The final output $\mathbf{O}$ is then forwarded as the output. The 'Flip' function is used to reverse the order of the sequence for the backward SSM and then restore the original order after processing. Element-wise summation is used to merge the forward and backward information flows, and the MLP serves to integrate across the channel dimension, which is empirically critical for performance.

4 Experiments

In this section, we introduce the incorporation of temporal SSM layers as temporal layers in diffusion models for video generation. This method facilitates the generation of longer video sequences while maintaining generative performance, overcoming the limitations imposed by memory constraints in attention-based temporal layers of video diffusion models. To empirically validate our hypothesis, we conducted a series of experiments comparing our temporal SSM layer with the temporal attention layer in video diffusion models. We also compared temporal linear attention, which has a linear spatial complexity similar to our temporal SSM layer.

4.1 Experimental Setup

Datasets We used two datasets to compare the performance of the temporal layers under varying conditions. Training details for each dataset and each frame length are shown in Appendix A. UCF101 (Soomro et al., 2012) was selected as the standard video dataset. Following Ho et al. (2022b), 13,320 videos from the training and test sets were used. We sampled clips with frame lengths of 16 and downscaled their spatial resolution to 32×32 or 64×64 pixels for manageability. Additionally, we incorporated the MineRL Navigate dataset (Guss et al., 2019; Saxena et al., 2021), which includes 961 videos for training purposes. Each video has 500 frames. This choice was motivated by our aim to validate the model's performance with longer video sequences. In our experiments, we explored including 64, 200 and 400 frames length, and downscaled the spatial resolution to 32×32 pixels.

Baseline We established our experimental baseline using VDMs (Ho et al., 2022b). Our analysis was meticulously designed to alter only the temporal attention layers in VDMs with our proposed temporal SSM layers. This strategy facilitated a direct comparison with the existing temporal layers.

Dataset	UCF101		MineRL		
# of Frames	16	16	64	200	400
Resolution	32×32	64×64	32×32	32×32	32×32
Attention	272.152	**618.005**	**1073.339**	**1032.514**	–
Linear Attention	285.995	670.776	1175.453	1170.834	1072.001
Our SSM	**226.447**	634.546	1132.982	1116.339	**972.306**

Table 2: Comparison of FVD in UCF101 and MineRL Navigate. The absence of values for generating 400 frames using the attention mechanism is attributed to the training on NVIDIA A100 exceeding the maximum memory capacity, rendering the experiment infeasible.

Evaluation Metrics In our validation process, we evaluated the sample quality of the videos generated by the trained models. To evaluate the quality of the generated videos, we employ the Fréchet Video Distance (FVD) metirc (Unterthiner et al., 2018), using an I3D network pretrained on the Kinetics-400 dataset (Carreira & Zisserman, 2017). FVD is a recognized standard for assessing the quality of generated videos (Ho et al., 2022b; Ge et al., 2022; Singer et al., 2022; Ho et al., 2022a; Harvey et al., 2022), where lower scores denote superior quality. For the UCF101 dataset, we evaluated FVD using all 13,320 videos and 10,000 generated samples. For the MineRL Navigate dataset, the calculation involves all 961 videos and 1,000 generated samples.

4.2 Benchmark Results

We initially focus on a standard video benchmark of UCF101 to compare generative performance (measured using FVD) across different temporal layers. The quantitative results are in Table 2. The qualitative results of this generation are presented in Figure 3. It is observed that the temporal SSM layer outperforms both temporal attention and temporal linear attention in generative performance in the 32×32 setting and performs competitively in the 64×64 setting.

We then utilize MineRL Navigate, a dataset with longer-term sequences, to observe the ability of temporal SSM layers to generate video sequences of lengths unattainable by temporal attention layers, while competitively maintaining generative performance. The transitions in these metrics as the video length increases are illustrated in Table 2. Additionally, qualitative results of the generated videos are shown in Figure 4.

While SSM exhibits competitive generative performance compared to temporal attention, it's noteworthy that temporal attention encounters memory errors at sequence lengths 400, a limitation not faced by temporal SSM. This highlights SSM's enhanced capability in handling longer sequences without compromising computational efficiency. Furthermore, SSM outperforms temporal linear attention in terms of generative performance. This finding underscores SSM's superiority in its ability to generate high-quality videos in longer video generation.

5 Conclusion

Our experimental findings demonstrate that incorporating SSM into the temporal layers of diffusion models for video generation offers superior video modeling in terms of memory efficiency for handling increased sequence lengths compared to traditional models employing temporal attention, while maintaining competitive generative quality. It is empirically shown that, while attention-based video generation diffusion models struggle to generate longer videos due to memory constraints, SSM-based video diffusion models are capable of producing such videos.

The results of this study indicate that the incorporation of SSMs can lead to the development of long-term video generation models that demand fewer memory resources. This has noteworthy implications for broadening the accessibility of cutting-edge research in video generation diffusion models. Even institutions with limited computational resources can engage in this advanced field, potentially expediting the pace of research and innovation in video generation.

REFERENCES

J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Shmuel Bar David, Itamar Zimerman, Eliya Nachmani, and Lior Wolf. Decision s4: Efficient sequence-based rl via state spaces layers. In *International Conference on Learning Representations*, 2023.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023.

J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

Tri Dao, Daniel Y Fu, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. "hungry hungry hippos: Towards language modeling with state space models". In *The International Conference on Learning Representations (ICLR)*, 2023.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Fei Deng, Junyeong Park, and Sungjin Ahn. Facing off world model backbones: Rnns, transformers, and s4. *arXiv:2307.02064*, 2023.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. *arXiv preprint arXiv:2204.03638*, 2022.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. In *International Conference on Machine Learning*, 2022.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

K. Gregor and F. Besse. Temporal difference variational auto-encoder. *arXiv preprint arXiv:1806.03107*, 2018.

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

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in Neural Information Processing Systems*, 33: 1474–1487, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983, 2022.

William H Guss, Brandon Houghton, Nicholay Topin, Phillip Wang, Cayden Codel, Manuela Veloso, and Ruslan Salakhutdinov. Minerl: A large-scale dataset of minecraft demonstrations. *arXiv preprint arXiv:1907.13440*, 2019.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *arXiv preprint arXiv:2205.11495*, 2022.

Kaiming He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pp. 6840–6851, 2020.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022b.

Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. In *ECCV*, 2022.

T. Kim, S. Ahn, and Y. Bengio. Variational temporal abstraction. In *Advances in Neural Information Processing Systems*, pp. 11566–11575, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *International Conference on Learning Representations*, 2013.

David M Knigge, David W Romero, Albert Gu, Efstratios Gavves, Erik J Bekkers, Jakub Mikolaj Tomczak, Mark Hoogendoorn, and Jan-jakob Sonke. Modelling long range dependencies in nd: From task-specific to a general purpose cnn. In *International Conference on Learning Representations*, 2023.

Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, and Y. Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. *arXiv preprint arXiv:2303.03982*, 2023.

Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. In *The Eleventh International Conference on Learning Representations*, 2023.

Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4nd: Modeling images and videos as multidimensional signals with state spaces. In *Advances in Neural Information Processing Systems*, 2022.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

M. Saito, E. Matsumoto, and S. Saito. Temporal generative adversarial nets with singular value clipping. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Vaibhav Saxena, Jimmy Ba, and Danijar Hafner. Clockwork variational autoencoders. *Advances in Neural Information Processing Systems*, 34, 2021.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv:2209.14792*, 2022.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. doi: 10.48550/arXiv.1212.0402.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021.

Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1526–1535, 2018.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. Selective structured state-spaces for long-form video understanding. In *CVPR*, 2023.

Junxiong Wang, Jing Nathan Yan, Albert Gu, and Alexander M Rush. Pretraining without attention. *arXiv preprint arXiv:2212.10544*, 2022.

Junxiong Wang et al. Mambabyte: Token-free selective state space model. *arXiv preprint arXiv:2401.13660*, 2024.

Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. *arXiv preprint arXiv:2311.18257*, 2023.

Wilson Yan et al. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

Linqi Zhou, Michael Poli, Winnie Xu, Stefano Massaroli, and Stefano Ermon. Deep latent state space models for time-series generation. In *International Conference on Machine Learning*, 2023.

Lianghui Zhu et al. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

A TRAINING DETAILS

To ensure a fair comparison of the modules extracting the temporal relationships under the same resolution settings, all configurations except for the temporal layers, were the same in our experiments. We used NVIDIA V100 $\times 4$ or NVIDIA A100 $\times 8$ (from a cloud provider). Detailed configuration is shown in Table 3.

Dataset	UCF101		MineRL		
# of Frames	16	16	64	200	400
Resolution	32×32	64×64	32×32	32×32	32×32
Base channel size	64	64	64	64	64
Channel multipliers	1, 2, 4, 8	1, 2, 4, 8	1, 2, 4, 8	1, 2, 4, 8	1, 2, 4, 8
Time embedding dimension	1024	1024	1024	1024	1024
Time embedding linears	2	2	2	2	2
# of attention heads (for attentions)	8	8	8	8	8
Dims of attention (for attentions)	64	64	64	64	64
SSM hidden dims (for SSMs)	512	512	512	512	512
MLP hidden dims (for SSMs)	512	512	512	512	512
Denoising timesteps (T)	256	1000	256	256	256
Loss type	L2 loss of ϵ	L2 loss of ϵ	L2 loss of ϵ	L2 loss of ϵ	L2 loss of ϵ
Training steps	92k	106k	174k	255k	246k
Optimizer	Adam	Adam	Adam	Adam	Adam
Training learning rate	0.0003	0.0001	0.0003	0.0003	0.00005
Train batch size	64	64	8	8	8
EMA decay	0.995	0.995	0.995	0.995	0.995
GPUs	V100 $\times 4$	A100 $\times 8$	V100 $\times 4$	A100 $\times 8$	A100 $\times 8$
Training Time	72 hours	120 hours	72 hours	100 hours	120 hours

Table 3: Experimental Setup for UCF101, MineRL Navigate datasets.

B ADDITIONAL RESULTS

In Figure 2, we present how training memory consumption and inference time vary with video sequence length for each temporal layer, using 32×32 resolution images. Training memory consumption data is based on a batch size of 8, while inference times reflect sample generation on a single NVIDIA A100 GPU with the number of diffusion time steps T fixed at 256. Notably, for the attention layer, experiments were capped at 275 frames instead of 400 due to memory limitations of the devices, which restrict training of attention-based video diffusion models to a maximum of approximately 275 frames. We also show qualitative results in Figure 3 and Figure 4.

C ABLATION STUDY

We conducted an ablation study to investigate the effectiveness of each component within our temporal SSM layer. Through this analysis, we aim to identify the critical elements when integrating an SSM into the temporal layer of video diffusion models.

C.1 ABLATION STUDY OF TEMPORAL SSM LAYERS

In the pursuit of temporal SSM layer in video generation diffusion models, our empirical findings indicate a considerable influence of the temporal layer's architecture on model performance. We conduct an ablation study to investigate the effects of each component of temporal SSM layer. Through this analysis, we aim to reveal which elements are instrumental to perform high generative capacity when employing an SSM in the temporal layer of video diffusion models.

Bidirectional SSM Switching from bidirectional SSM to (unidirectional) SSM leads to a significant reduction in generative performance (Table 4). This result show that the bidirectional usage of SSMs significantly improves the ability of temporal SSM layer to capture temporal relationships.

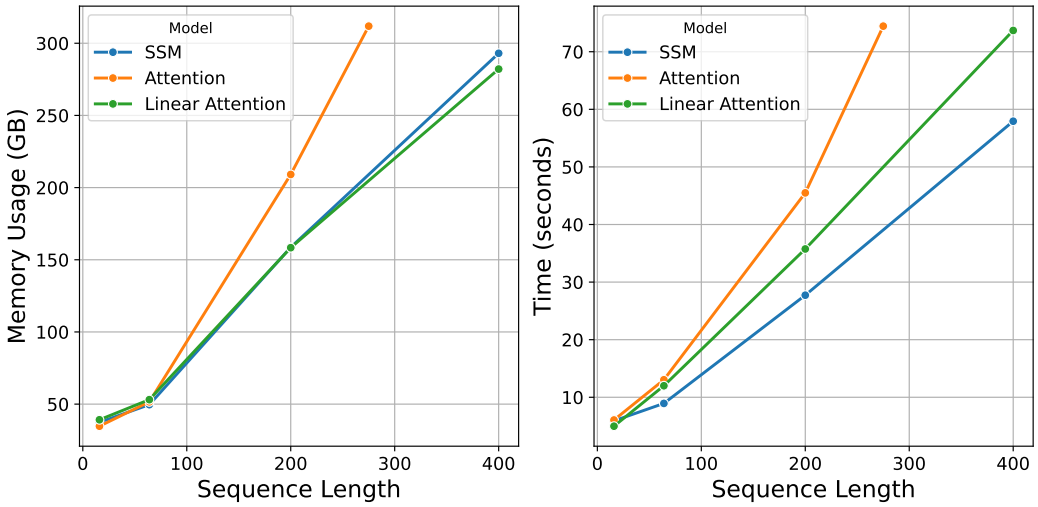


Figure 2: Left: Memory consumption during training with 8 NVIDIA A100 GPUs (40 GB) at a batch size of 8 and resolution of 32×32. Right: Inference time for generating a sample with a single NVIDIA A100 GPU at a resolution of 32×32 and $T = 256$.

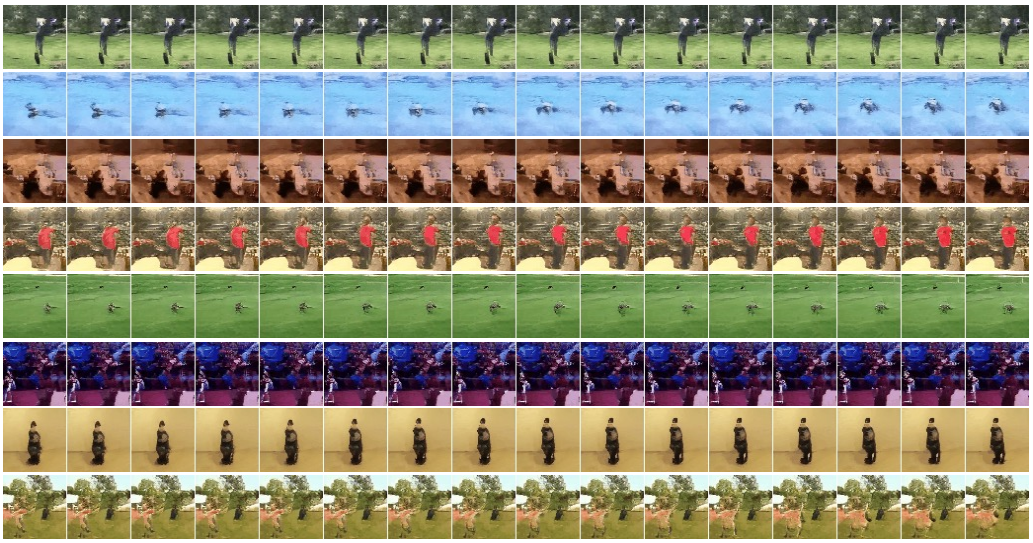


Figure 3: Qualitative generation results in UCF101 (64×64, number of frames are 16).

We aggregate features from Bidirectional SSMs through element-wise summation ("add" in Table4), unlike concatination-based aggregation approach taken by (Yan et al., 2023) ("concat" in the table). Through empirical validation, it has been demonstrated that this aggregation technique enhances the generative performance of video generation diffusion models.

MLPs in Temporal SSM Layers In this section, we shall discuss the configuration of the MLP within our temporal SSM layer. The proposed temporal SSM layer is composed of a bidirectional SSM followed by a dual-layer MLP with GeLU activation (Hendrycks & Gimpel, 2016). The rationale behind placing the MLP after the bidirectional SSM was that the SSM part of bidirectional SSM is a single-input, single-output system that was incapable of extracting relationships with other dimensions of the input. Therefore, the MLP was used to enhance the extraction of relationships with other input dimensions.

Figure 4: Qualitative results in MineRL Navigate (32×32, number of frames are 400).

SSM		MLP		FVD	
Bidirectional	Aggregation	# of Layers	Position	UCF101	MineRL-64
√	add	2	post	**226.447**	<u>1132.982</u>
√	add	2	pre	<u>253.688</u>	**1129.245**
	–	2	post	669.582	1371.131
√	concat	2	post	272.963	1224.561
√	add	1	post	267.935	1189.216
√	add	0	–	269.811	1193.491

Table 4: Quantitative results of ablation studies in our temporal SSM layers with UCF101 16 frames and MineRL 64 frames.

In this section, we describe the ablations preformed on the MLP layers. Specifically, we compared the following configurations: placing the MLP before the bidirectional SSM rather than after, replacing the MLP with a single layer and removing the MLP altogether. The results of the ablations are listed in Table 4.

Transitioning from a dual-layer MLP to a single linear layer, and eventually eliminating all such components, consistently leads to a stepwise degradation in generative performance across two datasets. The results also reveal that the placement of a dual-layer MLP, whether before or after the SSM, does not significantly affect the generative capability. These findings imply that in video diffusion models employing a temporal SSM layer, the generative performance heavily relies on the expressive capacity of the temporal SSM layer, while the positioning of components relative to the SSM plays a less crucial role in determining performance.

C.2 Comparison with Prior SSM Architectures

Recent years have seen a surge in research efforts to apply SSMs across various domains, accompanied by the proposal of numerous SSM architectures. Within this burgeoning field of study, gated state-space (GSS) architectures have emerged as one of the effective structures (Mehta et al., 2023). For instance, in Wang et al. (2022), bidirectional gated SSM (BiGS) architectures are employed for language modeling and image generation (Wang et al., 2022; Yan et al., 2023). GSS and BiGS architectures are composed of single linear layers for linear transformation, SSMs, skip connections implemented through multiplication (gating), and a two-layer MLP.

Among other effective SSM architectures, the Mamba architecture has been proposed (Gu & Dao, 2023). The Mamba architecture is characterized by a structure that alternates between 1D Convolution and SSM, and when combined with an SSM known as S6 (Gu & Dao, 2023), it has been applied to language understanding and learning of visual representations (Wang et al., 2024; Liu et al., 2024). For the learning of visual representations, a bidirectional version of Mamba, referred to here as

Architecture Settings			FVD	
Architectures	Bidirectional	Dims of MLPs	UCF101	MineRL-64
GSS		$d_{input} \times 3$	699.383	1683.820
BiGS	$\surd$	$d_{input} \times 3$	<u>233.277</u>	1157.793
BiGS	$\surd$	512	237.244	<u>1153.669</u>
Mamba		-	669.572	1722.097
Bi-Mamba	$\surd$	-	275.416	1156.891
Bi-Mamba + MLP	$\surd$	512	<u>243.638</u>	<u>1138.779</u>
Ours	$\surd$	512	**226.447**	**1132.982**

Table 5: FVD Comparison with SSM layers proposed in prior works across UCF101 16 frames and MineRL Navigate 64 frames. The resolution of both datasets is 32×32 pixels. The dimensionality of the input processed by the temporal layer, d_{input}, is equal to the channel size, C, as inferred from the dimensions of the input tensor $\mathbf{X} \in \mathbb{R}^{(B \times H \times W) \times L \times C}$. In the experimental settings described in this table, the channel sizes in the upper layers of the UNet architecture follow a sequential order of 64, 128, 256, and 512, starting from the topmost layer and progressing downwards.

Bi-Mamba, has also been proposed (Zhu et al., 2024). The architectures of these bidirectional SSM layers, which are more complex compared to our proposed temporal SSM layer.

It is important to note that in this section, with the goal of ensuring a fair comparison of architectures, the experiments with GSS, BiGS, Mamba, and Bi-Mamba all utilize the S4D as the SSM, similar to our temporal S4D layer. Architecture details are shown in Figure 5.

The comparisons between GSS and BiGS, as well as between Mamba and Bi-Mamba, suggest that the incorporation of bidirectionality significantly contributes to the capability of generating videos, a phenomenon not limited to our temporal SSM layers. Furthermore, introducing a Multilayer Perceptron (MLP) with a number of hidden layer dimensions that do not depend on the input size, following the SSM layer as done in our temporal SSM layers, to both BiGS and Bi-Mamba architectures, has been observed to result in performance improvements in some cases.

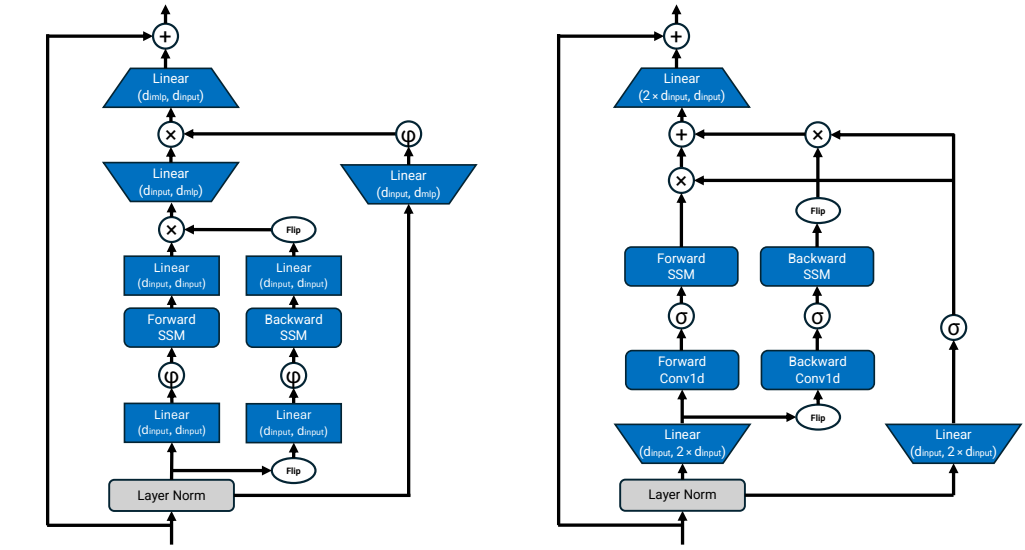
Then, we compare the generative performance of these structures and our proposed temporal SSM layer when used as the temporal layer in VDMs. The results indicate that, even when bidirectionality is applied and the size of the MLP's hidden layers is aligned with ours, these models exhibit inferior generative performance compared to our simple temporal SSM layer. A potential reason for this discrepancy is that while GSS, BiGS, Mamba and Bi-Mamba focus on language and image modeling using solely the SSM layers, our study seeks a temporal SSM layer specifically tailored to capture the temporal features of video, making the former approaches excessively redundant for this purpose.

D Related Works

D.1 Deep Generative Models for Video Generation

The field of video synthesis has seen significant advancements through various studies. Prior to the emergence of diffusion models, the use of generative adversarial networks (GANs) (Goodfellow et al., 2014) dominated the scene. These methods extended traditional image-GAN frameworks to video generation, focusing on enhancing their generative capabilities (Vondrick et al., 2016; Saito et al., 2017; Tulyakov et al., 2018; Ge et al., 2022). These approaches primarily aimed to achieve their objectives by extending common architectures of GANs for image generation. Additionally, the development of long-term video generation techniques, particularly those leveraging the transitions of latent variables in variational autoencoders (VAEs) (Kingma & Welling, 2013) are also well-known. (Kim et al., 2019; Gregor & Besse, 2018; Saxena et al., 2021; Yan et al., 2021).

The advent of diffusion models in image generation (Sohl-Dickstein et al., 2015; Nichol & Dhariwal, 2021; Ho et al., 2020) marked a turning point, with their subsequent application to video distributions demonstrating promising outcomes (Ho et al., 2022b; Singer et al., 2022; Ho et al., 2022a; Harvey et al., 2022; Blattmann et al., 2023). These approaches have shown promising results. Nevertheless,



(a) Gated State-Space (GSS), Bidirectional Gated SSM (BiGS) (b) Mamba, Bidirectional Mamba(Bi-Mamba)

Figure 5: Details of architectures in prior works (Mehta et al., 2023; Wang et al., 2022; Gu & Dao, 2023; Zhu et al., 2024). $\oplus$ means element-wise summation and $\otimes$ means element-wise product. ϕ means GeLU activation (Hendrycks & Gimpel, 2016) and σ means SiLU activation (Elfwing et al., 2018). In "Bi-Mamba + MLP" experiments in Section 5, we replaced the last linear layer in Bi-Mamba with a two-layer MLP.

the recent approaches adopt attention mechanisms to temporal layers, which requires memory proportional to the square of the sequence length, the computational and memory demands of video-diffusion models pose a substantial challenge. To mitigate this:

Spatiotemporal Downsampling Techniques like Singer et al. (2022) and Ho et al. (2022a) have been developed to reduce computational costs by lowering spatial resolution and temporal frequency while employing temporal attention layers for capturing temporal features. These methods complement our research, which concentrates on lightening the module that captures temporal dynamics.

Latent Diffusion Models Exploring latent diffusion models presents an alternative approach (Rombach et al., 2022), focusing on simpler latent variables rather than directly processing complex data. Some strategies involve using pre-trained image diffusion models with additional trainable layers for handling temporal features, emphasizing the training of these temporal layers to lessen the overall computational load (Blattmann et al., 2023).

Flexible Diffusion Models (FDMs) In the realm of long-term video prediction using video diffusion models, Flexible Diffusion Models (FDM) (Harvey et al., 2022) stand out. FDM predicts videos through autoregressive sampling with flexible frame conditioning, allowing for efficient long-term dependency modeling with minimal memory usage. Our research diverges from this by generating videos from scratch, rather than predicting based on initial frames, and by focusing on architectural improvements over sampling techniques.

D.2 Structured State Space Sequence (S4) Models

Originally introduced in Gu et al. (2021), S4 represents a sequence modeling framework that first solved all tasks in the Long-Range Arena (Tay et al., 2021). At its core lies the structured parameterization of state-space models (SSMs), offering efficient computation and demonstrating outstanding performance, in capturing long-range dependencies (Gu et al., 2020; 2021). The mathematical foundation of S4 is complex, prompting recent efforts to demystify, simplify, and improve S4 (Gu et al., 2022; Smith et al., 2023; Gu & Dao, 2023). The S4D (Gu et al., 2022) used in our study is one of such developments.

S4 and its variants have been applied across various domains, including image and video classification (Nguyen et al., 2022; Knigge et al., 2023; Islam & Bertasius, 2022; Wang et al., 2023), image representation learning (Liu et al., 2024), speech generation (Goel et al., 2022), time-series generation (Zhou et al., 2023), language modeling (Mehta et al., 2023; Wang et al., 2022; Dao et al., 2023), reinforcement learning (Bar David et al., 2023; Lu et al., 2023; Deng et al., 2023). In the field of diffusion models using SSMs, DiffuSSM (Yan et al., 2023) have explored the integration of SSMs with diffusion models, replacing the computationally intensive spatial attention mechanisms in image generation with SSMs. However, the application of SSMs in video generation diffusion models remains underexplored, an area our research aims to delve into and expand upon.

E EXPERIMENTAL CODES

Our codes are available at `https://github.com/shim0114/SSM-Meets-Video-Diffusion-Models`.