Matten: Video Generation with Mamba-Attention

Yu Gao[1], **Jiancheng Huang**[1], **Xiaopeng Sun**[1], **Zequn Jie**[†1], **Yujie Zhong**[1]
Lin Ma[1]
[1]Meituan Inc.

Abstract

In this paper, we introduce Matten, a cutting-edge latent diffusion model with Mamba-Attention architecture for video generation. With minimal computational cost, Matten employs spatial-temporal attention for local video content modeling and bidirectional Mamba for global video content modeling. Our comprehensive experimental evaluation demonstrates that Matten has competitive performance with the current Transformer-based and GAN-based models in benchmark performance, achieving superior FVD scores and efficiency. Additionally, we observe a direct positive correlation between the complexity of our designed model and the improvement in video quality, indicating the excellent scalability of Matten.

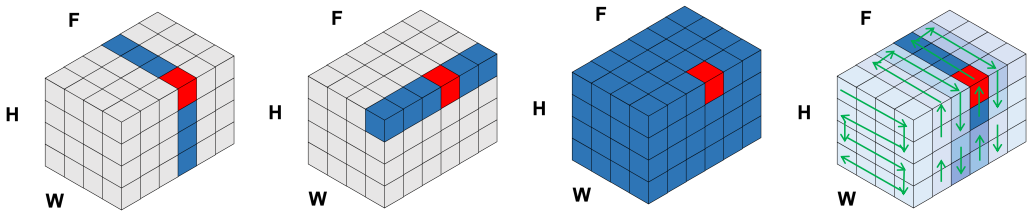
1 Introduction

Recent advancements in diffusion models have demonstrated impressive capabilities in video generation [1–5]. It has been observed that breakthroughs in architectural design are crucial for the efficient application of these models [6–8]. Contemporary studies largely concentrate on CNN-based U-Net architectures [1, 4] and Transformer-based frameworks [3, 2], both of which employ attention mechanisms to process spatio-temporal dynamics in video content. Spatial attention, which involves computing self-attention among image tokens within a single frame, is extensively utilized in both U-Net-based and Transformer-based video generation diffusion models as shown in Fig. 1 (a). Prevailing techniques typically apply local attention within the temporal layers as illustrated in Fig. 1 (b), where attention calculations are confined to identical positions across different frames. This approach fails to address the critical aspect of capturing interrelations across varying spatial positions in successive frames. A more effective method for temporal-spatial analysis would involve mapping interactions across disparate spatial and temporal locations, as depicted in Fig. 1 (c). Nonetheless, this global-attention method is computationally intensive due to the quadratic complexity involved in computing attention, thus requiring substantial computational resources.

There has been a rise in fascination with state space models (SSMs) across a variety of fields, largely due to their ability to deal with long sequences of data [9–11]. In the field of Natural Language Processing (NLP), innovations such as the Mamba model [10] have significantly improved both the efficiency of data inference processes and the overall performance of models by introducing dynamic parameters into the SSM structure and by building algorithms tailored for better hardware compatibility. The utility of the Mamba framework has been successfully extended beyond its initial applications, demonstrating its effectiveness in areas such as vision [12, 13] and multimodal applications [14]. Given the complexity of processing video data, we propose to use the Mamba architecture to explore spatio-temporal interactions in video content, as shown in Fig. 1 (d). However, unlike the self-attention layer, it's important to note that Mamba scans, which do not inherently compute dependencies between tokens, struggle to effectively detect localised data patterns, a limitation pointed out by [15].

[†] Corresponding to Zequn Jie <zequn.nus@gmail.com>.

Preprint. Under review.



(a) Spatial-Attention (b) Local Temporal-Attention (c) Global-Attention (d) Global-Mamba

Figure 1: Different ways of spatio-temporal modeling using Mamba and Attention. H, W, and F denote the height, weight, and frames, respectively. The red token is an example query, and the blue tokens mean those tokens having information interaction with the query. The shade of blue represents the intensity of the information interaction, with darker colors representing more direct interactions. Mamba scan interactions are distance-related between tokens with a linear complexity, while attention interactions are equal among these tokens with a quadratic complexity. For simplicity, we only show the unidirectional Mamba scan.

Regarding the advantages of Mamba and Attention, we introduce a latent diffusion model for video generation with a Mamba-Attention architecture, namely **Matten**. Specifically, we investigated the impact of various combinations of Mamba and Attention mechanisms on video generation. Our findings demonstrate that the most effective approach is to utilize the Mamba module to capture global temporal relationships (Fig. 1 (d)) while employing the Attention module for capturing spatial and local temporal relationships (Fig. 1 (a) and Fig. 1 (b)).

We conducted experimental evaluations to examine the performance and effects of Matten in both unconditional and conditional video generation tasks. Across all test benchmarks, Matten consistently exhibits the comparable FVD score [16] and efficiency with SOTAs. Furthermore, our results indicate that Matten is scalable, evidenced by the direct positive relationship between the model's complexity and the quality of generated samples.

In summary, our contributions are as follows:

- We propose Matten, a novel video latent diffusion model integrated with the mamba block and attention operations, which enables efficient and superior video generation.
- We design four model variants to explore the optimal combination of Mamba and attention in video generation. Based on these variants, we find that the most favorable approach is adopting attention mechanisms to capture local spatio-temporal details and utilizing the Mamba module to capture global information.
- Comprehensive evaluations show that our Matten achieves comparable performance to other models with lower computational and parameter requirements and exhibits strong scalability.

2 Related Work

2.1 Video Generation

The task of video generation primarily focuses on produce realistic video clips characterized by high-quality visuals and fluid movements. Previous video generation work can be grouped into 3 types. Initially, a number of researchers focused on adapting powerful GAN-based image generation techniques for video creation [17–21]. Nonetheless, GAN-based methods may lead to problems such as mode collapse, reducing diversity and realism.
In addition, certain models suggest the learning of data distributions via autoregressive models [22–25]. These methods typically yield high-quality videos and demonstrate more reliable convergence, but they are hindered by their substantial computational demands. Finally, the latest strides in video generation are centered on the development of systems that utilize diffusion models [26, 27, 4, 28–33, 2], which have shown considerable promise. These methods primarily use CNN-based U-Net or Transformer as the model architecture. Distinct from these works, our method concentrates on investigating the underexplored area of the combination of mamba and attention within video diffusion.

2.2 Mamba

Mamba, a new State-Space Model, has recently gained prominence in deep learning for its universal approximation capabilities and efficient modeling of long sequences, with applications in diverse fields such as medical imaging, image restoration, graphs, NLP, and image generation [34–40]. Drawing from control systems and leveraging HiPPO initialization [41], these models, like LSSL [11], address long-range dependencies but are limited by computational demands. To overcome this, S4 [42] and other structured state-space models introduce various configurations [43, 44, 9] and mechanisms [10] that have been integrated into larger representation models [45–47] for tasks in language and speech. Mamba, and its iterations like VisionMamba [12, 13], S4ND [48], and Mamba-ND [49], exhibit a range of computational strategies, from bidirectional SSMs to local convolution and multi-dimensionality considerations. For 3D imaging, T-Mamba [50] tackles the challenges in orthodontic diagnosis due to the powerful ability of Mmaba to handle long-range dependencies. For video understanding, VideoMamba [51] and Video Mamba Suite [52] adapt Mamba to the video domain and address the challenges of local redundancy and global dependencies prevalent in video data. In the domain of diffusion applications using mamba, Zigzag Mamba [53] advances the scalability and efficiency of generating visual content. It tackles the crucial problem of spatial continuity with an innovative scanning approach, incorporates text-conditioning features, and shows enhanced performance across high-resolution image and video datasets. [54] closely relates to our work, employing the mamba block in the temporal layer of video diffusion. Diverging from previous research focused mainly on local temporal modeling, our method, Matten, is uniquely designed to encompass global temporal dimensions.

3 Methodology

Our discussion starts with a brief overview of the latent space diffusion model and state space model in Sec. 3.1. This is followed by an in-depth description of the Matten model variants in Sec. 3.2. We then explore conditional ways related to timestep or class in Sec. 3.3. Lastly, a theoretical analysis comparing Mamba with Attention mechanisms is presented in Sec. 3.4.

3.1 Background

Latent Space Diffusion Models. [55]. For an input data sample $x \in p_{\text{data}}(x)$, Latent Diffusion Models (LDMs) initially utilize the pre-trained VAE or VQ-VAE encoder $\mathcal{E}$ to transform the data sample into a latent representation $z = \mathcal{E}(x)$. This transformation is followed by a learning phase where the data distribution is modeled through diffusion and denoising steps.

During the diffusion phase, noise is incrementally added to the latent encoding, producing a series of increasingly perturbed latent states z_t, where the intensity of additive noise is denoted by the timesteps $t \in T$. A specialized model such as U-Net ϵ_θ is utilized as the noise estimate network to estimate the noise perturbations affecting the latent representation z_t during the denoising phase, aiming to minimize the latent diffusion objective.

$$\mathcal{L}_{simple} = \mathbb{E}_{\mathbf{z}\sim p(z),\ \epsilon\sim\mathcal{N}(0,\mathbf{I}),\ t}\left[\|\epsilon - \epsilon_\theta(\mathbf{z}_t, t)\|_2^2\right]. \tag{1}$$

Furthermore, the diffusion models ϵ_θ are enhanced with a learned reverse process covariance Σ_θ, optimized using $\mathcal{L}_{vlb}$ as outlined by [6].

In our research, ϵ_θ is designed using a Mamba-based framework. Both $\mathcal{L}_{simple}$ and $\mathcal{L}_{vlb}$ are employed to refine the model's effectiveness and efficiency.

State Space Backbone. State space models (SSMs) have been rigorously validated both theoretically and through empirical evidence to adeptly manage long-range dependencies, demonstrating linear scaling with the length of data sequences. Conventionally, a linear state space model is represented as the following type:

$$\begin{aligned} h'(t) &= \mathbf{A}(t)h(t) + \mathbf{B}(t)x(t), \\ y(t) &= \mathbf{C}(t)h(t) + \mathbf{D}(t)x(t), \end{aligned} \tag{2}$$

which describes the transformation of a 1-D input sequence $x(t) \in \mathbb{R}$ into a 1-D output sequence $y(t) \in \mathbb{R}$, mediated by an N-D latent state sequence $h(t) \in \mathbb{R}^N$. State space models are particularly crafted to integrate multiple layers of these basic equations within a neural sequence modeling architecture, allowing the parameters $\mathbf{A}, \mathbf{B}, \mathbf{C}$, and $\mathbf{D}$ of each layer to be optimized via deep learning on loss function. N represents the state size, $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, and $\mathbf{D} \in \mathbb{R}$.

The process of discretization, essential for applying state space models as detailed in Eq. 2 to real-world deep learning tasks, converts continuous system parameters like $\mathbf{A}$ and $\mathbf{B}$ into their discrete

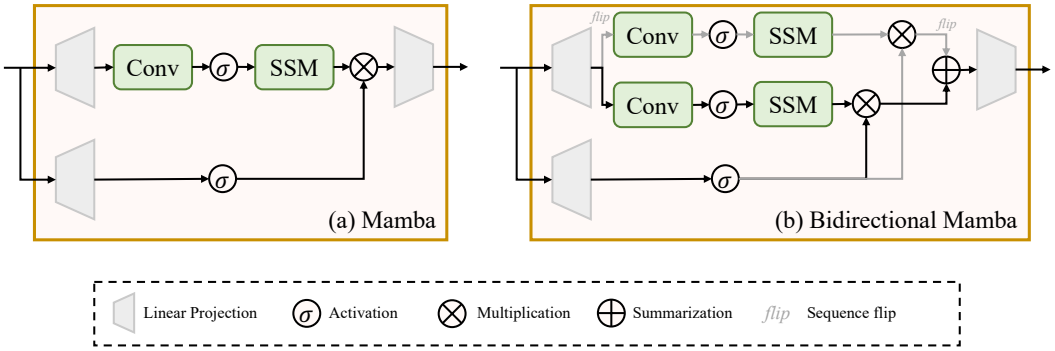


Figure 2: The original 1D sequence Mamba block and 2D bidirectional Mamba block. The normalization and the residual are omitted for simplification.

equivalents $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$. This critical step typically utilizes the zero-order hold (ZOH) method, a technique well-established in academic research for its efficacy. The ZOH method uses the timescale parameter Δ to bridge the gap between continuous and discrete parameters, thereby facilitating the application of theoretical models within computational settings.

$$\begin{aligned} \overline{\mathbf{A}} &= \exp(\Delta \mathbf{A}), \\ \overline{\mathbf{B}} &= (\Delta \mathbf{A})^{-1}(\exp(\mathbf{A}) - \mathbf{I}) \cdot \Delta \mathbf{B}. \end{aligned} \tag{3}$$

With these discretized parameters, the model outlined in Eq. 2 is then adapted to a discrete framework using a timestep Δ:

$$\begin{aligned} h_k &= \overline{\mathbf{A}} h_{k-1} + \overline{\mathbf{B}} x_k, \\ y_k &= \mathbf{C} h_k + \mathbf{D} x_k. \end{aligned} \tag{4}$$

This approach allows for the seamless integration of state space models into digital platforms. The traditional Mamba block, initially crafted for 1D sequence processing as shown in Fig. 2, is not ideally suited for visual tasks that demand spatial cognizance. To address this limitation, Vision Mamba [13] has developed a bidirectional Mamba block specifically tailored for vision-related applications. This innovative block is engineered to handle flattened visual sequences by employing both forward and backward SSMs concurrently, significantly improving its ability to process with spatial awareness. Mamba employs a work-efficient parallel scan that effectively reduces the sequential dependencies typically associated with recurrent computations. This optimization, coupled with the strategic utilization of GPU operations, eliminates the necessity to explicitly manage the expanded state matrix. In our study, we explore the integration of the Mamba architecture within a video generation framework, leveraging its efficiency and scalability.

3.2 The Model Variants of Matten

Consider the representation of a video clip's latent space, represented by $\boldsymbol{V_L} \in \mathbb{R}^{F \times H \times W \times C}$, where F indicates the number of frames, H the height of the frame, W the width of the frame, and C the channels per frame within the video's latent configuration. We transform $\boldsymbol{V_L}$ into a sequence of tokens by segmenting and reshaping it, represented as $\hat{\boldsymbol{z}} \in \mathbb{R}^{(n_f \times n_h \times n_w) \times d}$. Here, $n_f \times n_h \times n_w$ denotes the total number of tokens, with each token having dimension d.
Adopting a strategy similar to Latte, we assign $n_f = F$, $n_h = H/2$, and $n_w = W/2$ to structure the data effectively. Furthermore, a spatio-temporal positional embedding, denoted as $\boldsymbol{p}$, is incorporated into the token sequence $\hat{\boldsymbol{z}}$. The input for the Matten model thus becomes $\boldsymbol{z} = \hat{\boldsymbol{z}} + \boldsymbol{p}$, facilitating complex model interactions. As illustrated in Fig. 3, we introduce four distinct variants of the Matten model to enhance its versatility and effectiveness in video processing.
Global-Sequence Mamba Block. As illustrated in Fig. 3 (a), this variant refers to the execution of 3D Mamba scans in the full sequence of this spatiotemporal input. Following VideoMamba [51], we adopt *Spatial-First Scan* for our Global-Sequence Mamba block. This straightforward operation has been proven to be highly effective. It involves arranging spatial tokens based on their location and stacking them sequentially frame by frame. We reshape $\boldsymbol{z}$ into $\boldsymbol{z_{full}} \in \mathbb{R}^{1 \times n_f * n_h * n_w \times d}$ as the input of the Global-Sequence Mamba block to capture spatial-first information. The Bidirectional-Mamba layer is used.

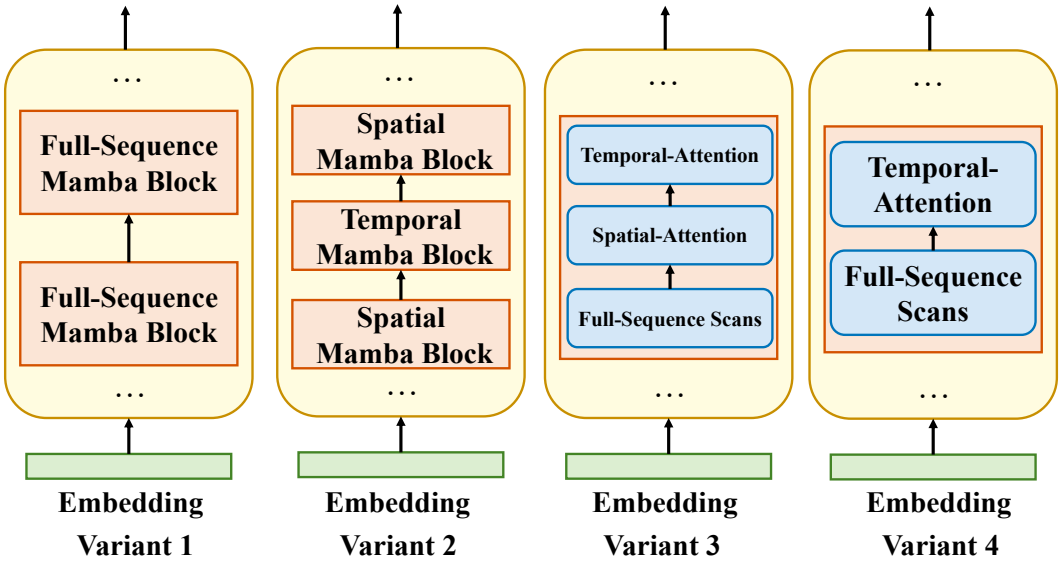


Figure 3: We introduce four model variants designed to harness spatio-temporal dynamics in videos effectively. For clarity, the embeddings shown in the diagram represent the patch and reshaped outcomes of the latent video.

Spatial and Temporal Mamba Blocks Interleaved. This particular variant leverages the Mamba module as a substitute for the traditional attention module within Transformer-based diffusion models for video generation, as noted in studies such as [2, 56, 57]. Illustrated in Fig. 3 (b), the backbone of this variant, known as Matten, is equipped with two types of Bidirectional-Mamba blocks: spatial Bidirectional-Mamba blocks and temporal Bidirectional-Mamba blocks. The spatial blocks are designed to solely capture spatial details among tokens that share identical temporal indices, whereas the temporal blocks are tasked with capturing information across different times within the same spatial coordinate. For effective spatial information processing, $\boldsymbol{z}$ is restructured into $\boldsymbol{z_s} \in \mathbb{R}^{n_f \times s \times d}$, which then serves as the input for the spatial Mamba block.
Then, we reshape $\boldsymbol{z_s}$ into $\boldsymbol{z_t} \in \mathbb{R}^{s \times n_f \times d}$ for the temporal Mamba block to process temporal information.
Global-Sequence Mamba Block with Spatial-Temporal Attention Interleaved. Although Mamba demonstrates efficient performance in long-distance modeling, its advantages in shorter sequences modeling are not as pronounced [10], compared to the attention operation in Transformer. Consequently, we have developed a hybrid block that leverages the strengths of both the attention mechanism and Mamba as illustrated in Fig. 3 (c), which integrates Mamba and Attention computations for both short and long-range modeling. Each block is composed of Spatial Attention computation, Temporal Attention computation, and a Global-Sequence Mamba scan in series. This design enables our model to effectively capture both the global and local information present in the latent space of videos.
Global-Sequence Mamba Block with Temporal Attention Interleaved.
The scanning in the Global-Sequence Mamba block is continuous in the spatial domain but discontinuous in the temporal domain [51]. Thus, this variant has removed the Spatial Attention component, while retaining the Temporal Attention block. Consequently, by concentrating on a Spatial-First scan augmented with Temporal Attention shown in Fig. 3 (d), we strive to enhance our model's efficiency and precision in processing the dynamic facets of video data, thereby assuring robust performance in a diverse range of video processing tasks.

3.3 Conditional Way of Timestep or Class

Drawing from the frameworks presented by Latte and DiS, we perform experiments on two distinct methodologies for embedding timestep or class information c into our model. The first method, inspired by DiS, involves treating c as tokens, a strategy we designate as *conditional tokens*. The second method adopts a technique akin to adaptive normalization (AdaN) [58, 7], specifically tailored for integration within the Mamba block. This involves using MLP layer to compute parameters

Method	Pretrained	FaceForensics	SkyTimelapse	UCF101	Taichi-HD	FLOPs (G)
MoCoGAN	✗	124.7	206.6	2886.9	-	
VideoGPT	✗	185.9	222.7	2880.6	-	
DIGAN	✗	62.5	83.11	1630.2	156.7	
StyleGAN-V	✗	47.41	79.52	1431.0	-	
PVDM	✗	355.92	75.48	1141.9	540.2	
MoStGAN-V	✗	39.70	65.30	1380.3	-	
MoCoGAN-HD	✓	111.8	164.1	1729.6	128.1	
LVDM	✓	-	95.20	372.0	**99.0**	
Latte	✓	**34.00**	59.82	477.97	159.60	5572
Matten (ours)	✗	45.01	**53.56**	**210.61**	158.56	4008

Table 1: FVD metrics for various video generation models across multiple datasets are presented. FVD scores for comparative baseline models, as reported in sources such as Latte, StyleGAN-V, or respective original publications, are included for reference. In this context, "Pretrained" refers to models that utilize a pretraining approach based on image generation techniques.

γ_c and β_c from c, formulating the operation $AdaN(f, c) = \gamma_c \text{Norm}(f) + \beta_c$, where f denotes the feature maps in the Mamba block. Further, this adaptive normalization is implemented prior to residual connections of the Mamba block, implementing by the transformation $RCs(f, c) = \alpha_c f + MambaScans(AdaN(f, c))$, with $MambaScans$ representing the Bidirectional-Mamba scans within the block. We refer to this advanced technique as Mamba adaptive normalization (*M-AdaN*), which seamlessly incorporates class or timestep information to enhance model responsiveness and contextual relevance.

3.4 Analysis of Mamba and Attention

In summary, the hyperparameters of our proposed block encompass hidden size D, expanded state dimension E, and SSM dimension N. All the settings of Matten are detailed in Table 2, covering different numbers of parameters and computation cost to thoroughly evaluate scalability performance. Specifically, the Gflop metric is analyzed during the generation of 16×256×256 unconditional videos, employing a patch size of $p = 2$. Consistent with [10], we standardize the SSM dimension N across all models at 16.

Both the SSM block within Matten and the self-attention mechanism in Transformer architectures are integral for effective context modeling. We provide a detailed theoretical analysis of computational efficiency as well. For a given sequence $\mathbf{X} \in \mathbb{R}^{1\times J\times D}$ with the standard setting $E = 2$, the computational complexities of self-attention (SA), Feed-Forward Net (FFN) and SSM operations are calculated as follows:

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

$$\mathcal{O}(\text{FFN}) = 4JD^2, \tag{6}$$

$$\mathcal{O}(\text{SSM}) = 3J(2D)N + J(2D)N^2. \tag{7}$$

$3J(2D)N$ involves the calculation with $\overline{\mathbf{B}}$, $\mathbf{C}$, and $\mathbf{D}$, while $J(2D)N^2$ denotes the calculation with $\overline{\mathbf{A}}$. It demonstrates that self-attention's computational demand scales quadratically with the sequence length J, whereas SSM operations scale linearly. Notably, with N typically fixed at 16, this linear scalability renders the Mamba architecture particularly apt for handling extensive sequences typical in scenarios like global relationship modeling in video data. When comparing the terms $2J^2D$ and $J(2D)N^2$, it is clear that the Mamba block is more computationally efficient than self-attention, particularly when the sequence length J significantly exceeds N^2. For shorter sequences that focus on spatial and localized temporal relationships, the attention mechanism offers a more computationally efficient alternative when the computational overhead is manageable, as corroborated by empirical results.

4 Experiments

This part first describes the experimental settings, including details about the datasets we used, evaluation metrics, compared methods, configurations of the Matten model, and specific implementation

	Variant 1	Variant 2	Variant 3	Variant 4
Params (M)	814	814	853	846
FLOPs (G)	1590	1660	4008	3660

Table 2: The parameter count and FLOPs (Floating-Point Operations) associated with various model variants of Matten.

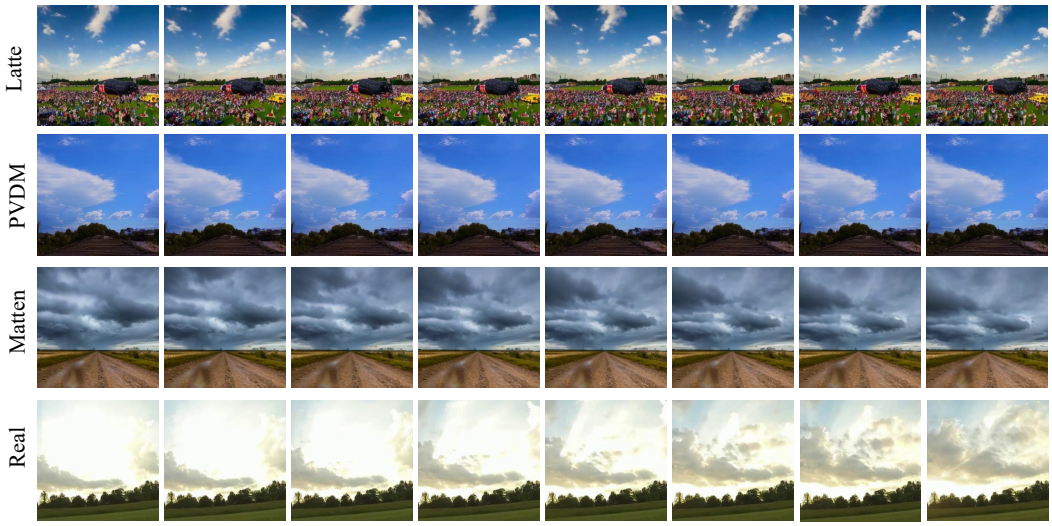


Figure 4: Sample videos from the different methods and real data on SkyTimelapse.

aspects. Following this, ablation studies are conducted to identify optimal practices and assess the impact of model size. The section concludes with a comparative analysis of our results on 4 common datasets against advanced video generation methods.

4.1 Experimental Detail

Datasets Overview. We engage in extensive experiments across four renowned and common datasets: FaceForensics [59], SkyTimelapse [60], UCF101 [61], and Taichi-HD [62]. Following protocols established in Latte, we utilize predefined training and testing divisions. From these datasets, we extract video clips consisting of 16 frames, applying a sampling interval of 3, and resize each frame to a uniform resolution of 256x256 for our experiments.

Evaluation Metrics. For robust quantitative analysis, we adopt the Fréchet Video Distance (FVD) [16], recognized for its correlation with human perceptual evaluation. In compliance with the methodologies of StyleGAN-V, we determine FVD scores by examining 2,048 video clips, each containing 16 frames.

Baseline Comparisons. Our study includes comparisons with advanced methods to assess the performance of our approach quantitatively, including MoCoGAN [63], VideoGPT [25], MoCoGAN-HD [64], DIGAN [65], StyleGAN-V [66], PVDM [1], MoStGAN-V [67], LVDM [68], and Latte [2]. Unless explicitly stated otherwise, all presented values are obtained from the latest relevant studies: Latte, StyleGAN-V, PVDM, or the original paper.

Matten Model Configurations. Our Matten model is structured using a series of L Mamba blocks, with each block having a hidden dimension of D. Inspired by the Vision Transformer (ViT) approach, we delineate four distinct configurations varying in parameter count, detailed in Table 3.

Implementation Specifics. All ablation experiments adopt the AdamW optimizer, set at a fixed learning rate of 1×10^{-4}. The sole augmentation technique applied is horizontal flipping. Consistent with prevailing strategies in generative modeling [7, 8], we employ the exponential moving average (EMA) of the model weights with a decay rate of 0.99 at the first 50k steps and the other 100k steps during the training process. The results reported are derived directly using the EMA-enhanced models. Additionally, the architecture benefits from the integration of a pre-trained variational autoencoder, sourced from Stable Diffusion v1-4.

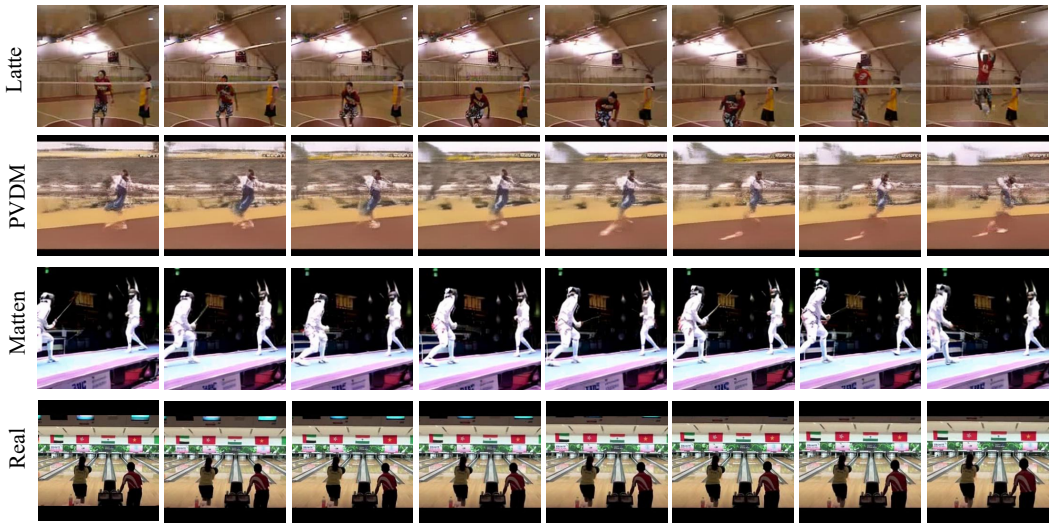


Figure 5: Sample videos generated using various methods on the UCF101 dataset, highlighting the visually appealing nature of the results.

Figure 6: Sample videos generated using various methods on the FaceForensics dataset, highlighting the visually appealing nature of the results.

4.2 Ablation study

In this part, we detail our experimental investigations using the SkyTimelapse dataset to assess the impact of various design modifications, model variations, and model sizes on performance, as previously introduced in Secs. 3.3 and 3.2.

Timestep-Class Information Injection Illustrated in Fig. 8b, the *M-AdaN* approach markedly outperforms *conditional tokens*. We surmise this difference stems from the method of integration of timestep or class information. *Conditional tokens* are introduced directly into the model's input, potentially creating a spatial disconnect within the Mamba scans. In contrast, *M-AdaN* embeds both timestep and class data more cohesively, ensuring uniform dissemination across all video tokens, and enhancing the overall synchronization within the model.

Exploring Model Variants Our analysis of Matten's model variants, as detailed in Sec. 3.2, aims to maintain consistency in parameter counts to ensure equitable comparisons. Each variant is developed from the ground up. As depicted in Fig. 8a, Variant 3 demonstrates superior performance with increasing iterations, indicating its robustness. Conversely, Variants 1 and 2, which focus primarily

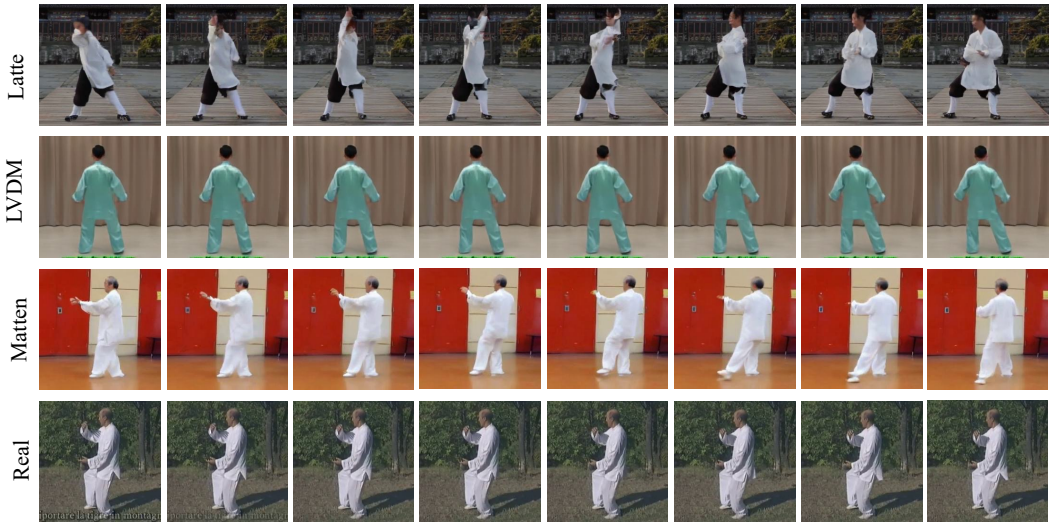
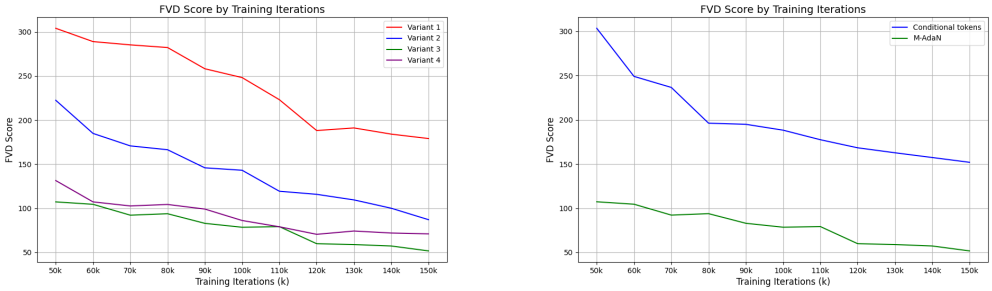


Figure 7: Sample videos generated using various methods on the Taichi-HD dataset, highlighting the visually appealing nature of the results.



(a) Model variants

(b) Timestep-class conditional way

Figure 8: Exploration of Design Choices Through Ablation Studies. We have conducted various ablation studies to identify optimal strategies for Mamba-based video diffusion models, focusing on improving FVD metrics on the SkyTimelapse dataset. For enhanced clarity, please magnify the displayed results.

on local or global information, respectively, lag in performance, underscoring the necessity for a balanced approach in model design.

Assessment of Model Size We experiment with four distinct sizes of the Matten model—XL, L, B, and S as listed in Tab. 3 on the SkyTimelapse dataset. The progression of their Fréchet Video Distances (FVDs) with training iterations is captured in Fig. 9. There is a clear trend showing that larger models tend to deliver improved performance, echoing findings from other studies in image and video generation [7], which highlight the benefits of scaling up model dimensions.

4.3 Comparison Experiment

According to the findings from the ablation studies presented in Sec. 4.2, we have pinpointed the settings about how to design our Matten, notably highlighting the efficacy of model variant 3 equipped with *M-AdaN*. Leveraging these established best practices, we proceed to conduct comparisons against contemporary state-of-the-art techniques.

Qualitative Assessment of Results Figures 4 through 7 display the outcomes of video synthesis using various methods across datasets such as UCF101, Taichi-HD, FaceForensics, and SkyTimelapse. Across these different contexts, our method consistently delivers realistic video generations at a high resolution of 256x256 pixels. Notable achievements include accurately capturing facial motions and effectively handling dynamic movements of athletes. Our model particularly excels in generating

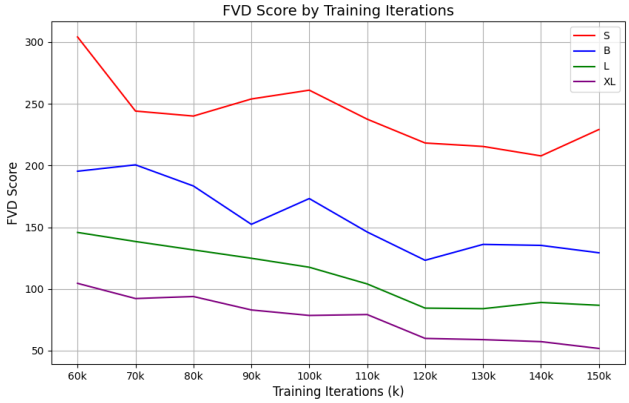


Figure 9: The impact of varying model sizes on performance is notable. Generally, enlarging the model dimensions tends to markedly enhance its effectiveness.

Model	Layer numbers L	Hidden size D	SSM dimension N	Param
Matten-S	12	384	16	35M
Matten-B	12	768	16	164M
Matten-L	24	1024	16	579M
Matten-XL	28	1152	16	853M

Table 3: Specifics of our model configurations adhere to the setups outlined for various model sizes following the ViT and DiT frameworks.

high-quality videos on the UCF101 dataset, an area where many other models frequently falter. This capability underscores our method's robustness in tackling complex video synthesis challenges.
Quantitative results. Tab. 1 presents the quantitative results of each comparative method. Overall, our method surpasses prior works and matches the performance of methods with image-pretrained weights, demonstrating our method's superiority in video generation. Furthermore, our model attains roughly a 25% reduction in flops compared to Latte, the latest Transformer-based model. Given the abundance of released pre-trained U-Net-based (Stable Diffusion, SDXL) or Transformer-based (DiT, PixArt) image generation models, these U-Net-based or Transformer-based video generation models can leverage these pre-trained models for training. However, there are no released, pre-trained Mamba-based image generation models yet, so our model has to be trained from scratch. We believe that once Mamba-based image generation models become available, they will be of great help in training our Matten.

5 Conclusion

This paper proposes a simple diffusion method for video generation, Matten, with the Mamba-Attention structure as the backbone for generating videos. To explore the quality of Mamba for generating videos, we explore different configurations of the model, including four model variants, time step and category information injection, and model size. Extensive experiments demonstrate that Matten excels in four standard video generation benchmarks and displays impressive scalability.

References

[1] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023.

[2] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[3] Haoyu Lu, Guoxing Yang, Nanyi Fei, Yuqi Huo, Zhiwu Lu, Ping Luo, and Mingyu Ding. Vdt: General-purpose video diffusion transformers via mask modeling. In *ICLR*, 2023.

[4] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *NeurIPS*, 2022.

[5] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *NeurIPS*, 2022.

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

[7] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, pages 4195–4205, 2023.

[8] Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. All are worth words: A vit backbone for diffusion models. In *Computer Vision and Pattern Recognition*, pages 22669–22679, 2023.

[9] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022.

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

[11] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *NeurIPS*, 2021.

[12] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

[13] Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. *arXiv preprint arXiv:2401.09417*, 2024.

[14] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. *arXiv preprint arXiv:2403.14520*, 2024.

[15] Simiao Zuo, Xiaodong Liu, Jian Jiao, Denis Charles, Eren Manavoglu, Tuo Zhao, and Jianfeng Gao. Efficient long sequence modeling via state space augmented transformer. *arXiv preprint arXiv:2212.08136*, 2022.

[16] 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.

[17] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Neural Information Processing Systems*, 29, 2016.

[18] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *International Conference on Computer Vision*, pages 2830–2839, 2017.

[19] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *Winter Conference on Applications of Computer Vision*, 2020.

[20] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. G3an: Disentangling appearance and motion for video generation. In *Computer Vision and Pattern Recognition*, pages 5264–5273, 2020.

[21] Emmanuel Kahembwe and Subramanian Ramamoorthy. Lower dimensional kernels for video discriminators. *Neural Networks*, 132:506–520, 2020.

[22] 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. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022.

[23] Ruslan Rakhimov, Denis Volkhonskiy, Alexey Artemov, Denis Zorin, and Evgeny Burnaev. Latent video transformer. In *Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2021.

[24] Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. In *International Conference on Learning Representations*, 2020.

[25] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Neural Information Processing Systems*, 33:6840–6851, 2020.

[27] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Neural Information Processing Systems*, 35:27953–27965, 2022.

[28] 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 preprint arXiv:2209.14792*, 2022.

[29] Kangfu Mei and Vishal Patel. Vidm: Video implicit diffusion models. In *AAAI Conference on Artificial Intelligence*, pages 9117–9125, 2023.

[30] 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 *Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[31] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.

[32] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023.

[33] Yaohui Wang, Xin Ma, Xinyuan Chen, Antitza Dantcheva, Bo Dai, and Yu Qiao. Leo: Generative latent image animator for human video synthesis. *arXiv preprint arXiv:2305.03989*, 2023.

[34] Jun Ma, Feifei Li, and Bo Wang. U-mamba: Enhancing long-range dependency for biomedical image segmentation. *arXiv preprint arXiv:2401.04722*, 2024.

[35] Junxiong Wang, Tushaar Gangavarapu, Jing Nathan Yan, and Alexander M Rush. Mambabyte: Token-free selective state space model. *arXiv preprint arXiv:2401.13660*, 2024.

[36] Zhuoran Zheng and Chen Wu. U-shaped vision mamba for single image dehazing. *arXiv preprint arXiv:2402.04139*, 2024.

[37] Dingkang Liang, Xin Zhou, Xinyu Wang, Xingkui Zhu, Wei Xu, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *arXiv preprint arXiv:2402.10739*, 2024.

[38] Zhengcong Fei, Mingyuan Fan, Changqian Yu, and Junshi Huang. Scalable diffusion models with state space backbone. *arXiv preprint arXiv:2402.05608*, 2024.

[39] Ali Behrouz and Farnoosh Hashemi. Graph mamba: Towards learning on graphs with state space models. *arXiv preprint arXiv:2402.08678*, 2024.

[40] Md Atik Ahamed and Qiang Cheng. Mambatab: A simple yet effective approach for handling tabular data. *arXiv preprint arXiv:2401.08867*, 2024.

[41] 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.

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

[43] Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems*, 35:22982–22994, 2022.

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

[45] Harsh Mehta, Ankit Gupta, Ashok Cutkosky, and Behnam Neyshabur. Long range language modeling via gated state spaces. *arXiv preprint arXiv:2206.13947*, 2022.

[46] Xuezhe Ma, Chunting Zhou, Xiang Kong, Junxian He, Liangke Gui, Graham Neubig, Jonathan May, and Luke Zettlemoyer. Mega: moving average equipped gated attention. *arXiv preprint arXiv:2209.10655*, 2022.

[47] Daniel Y Fu, Tri Dao, Khaled K Saab, Armin W Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. *arXiv preprint arXiv:2212.14052*, 2022.

[48] 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. *Advances in neural information processing systems*, 35:2846–2861, 2022.

[49] Shufan Li, Harkanwar Singh, and Aditya Grover. Mamba-nd: Selective state space modeling for multi-dimensional data. *arXiv preprint arXiv:2402.05892*, 2024.

[50] Jing Hao, Lei He, and Kuo Feng Hung. T-mamba: Frequency-enhanced gated long-range dependency for tooth 3d cbct segmentation. *arXiv preprint arXiv:2404.01065*, 2024.

[51] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. *arXiv preprint arXiv:2403.06977*, 2024.

[52] Guo Chen, Yifei Huang, Jilan Xu, Baoqi Pei, Zhe Chen, Zhiqi Li, Jiahao Wang, Kunchang Li, Tong Lu, and Limin Wang. Video mamba suite: State space model as a versatile alternative for video understanding. *arXiv preprint arXiv:2403.09626*, 2024.

[53] Vincent Tao Hu, Stefan Andreas Baumann, Ming Gui, Olga Grebenkova, Pingchuan Ma, Johannes Fischer, and Bjorn Ommer. Zigma: Zigzag mamba diffusion model. *arXiv preprint arXiv:2403.13802*, 2024.

[54] Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki, and Yutaka Matsuo. Ssm meets video diffusion models: Efficient video generation with structured state spaces. *arXiv preprint arXiv:2403.07711*, 2024.

[55] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[56] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

[57] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[58] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Antelligence*, 2018.

[59] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics: A large-scale video dataset for forgery detection in human faces. *arXiv preprint arXiv:1803.09179*, 2018.

[60] Wei Xiong, Wenhan Luo, Lin Ma, Wei Liu, and Jiebo Luo. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. In *Computer Vision and Pattern Recognition*, pages 2364–2373, 2018.

[61] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11), 2012.

[62] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. *Neural Information Processing Systems*, 32, 2019.

[63] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *Computer Vision and Pattern Recognition*, pages 1526–1535, 2018.

[64] Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. In *International Conference on Learning Representations*, 2021.

[65] Sihyun Yu, Jihoon Tack, Sangwoo Mo, Hyunsu Kim, Junho Kim, Jung-Woo Ha, and Jinwoo Shin. Generating videos with dynamics-aware implicit generative adversarial networks. In *International Conference on Learning Representations*, 2022.

[66] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Computer Vision and Pattern Recognition*, pages 3626–3636, 2022.

[67] Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Mostgan-v: Video generation with temporal motion styles. In *Computer Vision and Pattern Recognition*, pages 5652–5661, 2023.

[68] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2023.