Long-Context State-Space Video World Models

Ryan Po[1] Yotam Nitzan[3] Richard Zhang[3] Berlin Chen[2] Tri Dao[2]
Eli Shechtman[3] Gordon Wetzstein[1] Xun Huang[3]
[1]Stanford University [2]Princeton University [3]Adobe Research

Abstract

Video diffusion models have recently shown promise for world modeling through autoregressive frame prediction conditioned on actions. However, they struggle to maintain long-term memory due to the high computational cost associated with processing extended sequences in attention layers. To overcome this limitation, we propose a novel architecture leveraging state-space models (SSMs) to extend temporal memory without compromising computational efficiency. Unlike previous approaches that retrofit SSMs for non-causal vision tasks, our method fully exploits the inherent advantages of SSMs in causal sequence modeling. Central to our design is a block-wise SSM scanning scheme, which strategically trades off spatial consistency for extended temporal memory, combined with dense local attention to ensure coherence between consecutive frames. We evaluate the long-term memory capabilities of our model through spatial retrieval and reasoning tasks over extended horizons. Experiments on Memory Maze and Minecraft datasets demonstrate that our approach surpasses baselines in preserving long-range memory, while maintaining practical inference speeds suitable for interactive applications.

1. Introduction

World models [2, 23, 25, 31, 47, 70, 74, 92] are causal generative models designed to predict how world states evolve in response to actions, enabling interactive simulation of complex environments. Video diffusion models [7, 29, 37, 41, 45, 53, 85] have emerged as a promising approach for world modeling. While early models were limited to generating fixed-length videos and therefore unsuitable for interactive applications, recent architectures have enabled infinite-length video generation through autoregressive, sliding-window prediction [1, 9, 14, 16, 18, 26, 32, 36, 50, 60, 66, 73, 87, 88]. This paves the way for a new paradigm in which video diffusion models can interactively simulate visual worlds by continuously generating video frames conditioned on interactive control signals.

However, existing video world models have very limited temporal memory due to the restricted context length

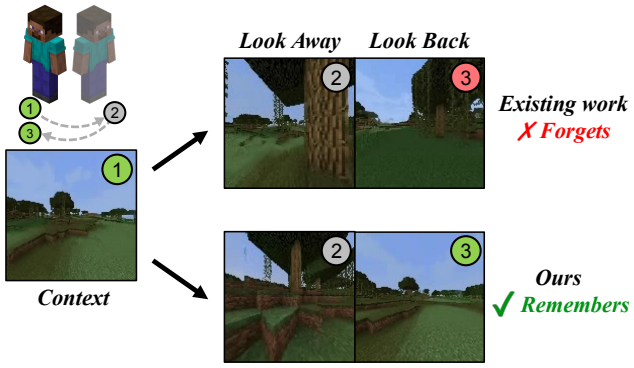

Figure 1. **Failure case of existing video world models.** Without long-term memory, previously observed regions may appear altered or inconsistent upon revisiting.

of their attention mechanisms. This limitation hinders their ability to simulate a persistent world with long-term consistency. For example, when using an existing video world model to simulate a game, the entire environment might completely change after a player simply looks right and then left again (see Fig. 1). This is because the frames that contain the original environment are no longer in the model's attention window. While one could theoretically extend the memory by training the model on longer context windows, this approach faces two major limitations: (1) the computational cost of training scales quadratically with context length, making it prohibitively expensive, and (2) the per-frame inference time grows linearly with context length, resulting in increasingly slower generation speed that would be impractical for applications that require real-time, infinite-length generation (such as gaming).

In this work, we propose a novel video world model architecture that leverages state-space models (SSMs) to enable long-term memory while maintaining efficient training and fast autoregressive inference. Our key innovation is a block-wise scan scheme of Mamba [20] that optimally balances temporal memory and spatial coherence, while preserving temporal causality. We complement this with dense local attention between neighboring frames, maintaining high-fidelity generation with minimal computational overhead. Unlike previous methods that retrofit SSMs for non-causal vision tasks, our approach fundamentally differs by

Architecture	Training Complexity	AR Inference (in total)	AR Inference (per frame)	Long Memory
Bidirectional attention	Quadratic	Cubic	Quadratic	✓
Causal attention	Quadratic	Quadratic	Linear	✓
Causal + sliding window inference	Sub-quadratic	Linear	Constant	✗
Ours (SSM + local causal attention)	Linear	Linear	Constant	✓

Table 1. Computational complexity comparison for autoregressive video diffusion architectures, with respect to sequence length. Bidirectional attention models process all previous frames with quadratic complexity when generating a single frame, while causal attention models improve efficiency through KV-caching but still scale linearly with video length. Both approaches commonly employ sliding-window inference to maintain manageable computational complexity. While sliding window inference with causal attention achieves constant per-frame inference time, it sacrifices long-term memory. Our hybrid architecture combines SSMs with local causal attention, maintaining long-term memory while achieving constant per-frame inference speed—ideal for interactively generating a persistent world.

specifically employing SSMs to handle causal temporal dynamics and track the state of the world, fully exploiting their inherent strengths for sequence modeling. Tab. 1 compares our method with existing solutions. Unlike architectures with full bidirectional [9, 60, 88] or causal attention [14, 16, 18, 32, 36, 50, 87], our model achieves constant per-frame inference time, making it particularly suitable for interactive applications that require infinite rollout. Furthermore, our approach maintains consistent long-term memory, unlike existing methods that apply causal attention within a sliding window [14, 87].

The remainder of this paper is structured as follows: Our approach builds on autoregressive video diffusion models trained with independent per-frame noise levels [10], as well as SSM architectures. Sec. 3 provides the necessary technical background on these components. In Sec. 4, we detail our novel architecture design that combines block-wise SSM scanning with local causal attention, along with our specialized training and sampling strategies. Sec. 5 introduces evaluation metrics that are used to assess long-term memory in video world models. Specifically, the spatial retrieval metric quantifies the model's ability to retrieve the environment when revisiting a previously observed location, while the spatial reasoning metric requires the model to infer the appearance at unvisited locations based on past observations from different viewpoints. Experiments demonstrate that our approach significantly outperforms baseline architectures under these metrics in challenging Memory Maze and Minecraft datasets, confirming its effectiveness in maintaining long-term memory.

2. Related Work

Video generation. Modern video generative models rely on either autoregressive (AR) prediction of discretized tokens [40, 63, 71, 78] or diffusion models. Compared to discretized AR models, diffusion models usually produce higher-quality videos. They first demonstrate remarkable success in image synthesis [54, 56] and have been extended to videos by treating the temporal dimension analogously to spatial dimensions, generating the entire video of a fixed length in a single process [6, 24, 29, 41, 42, 49, 58, 84].

Although effective for short clips, diffusion models face significant limitations when scaling to longer video sequences due to prohibitive computational demands. Additionally, they lack support for online, incremental generation of videos. To address this, researchers have explored training strategies for video diffusion models that enable AR inference. Some works train conditional diffusion models to denoise a few next frames conditioned on past clean frames [11, 17, 18, 27, 29, 36, 91], while others introduce per-frame independent noise levels during training which also enable AR inference [10, 60, 87]. At inference time, both approaches can generate long videos autoregressively, typically through a sliding-window mechanism due to computational constraints. This inherently restricts long-term memory and prevents the model from maintaining temporal consistency over extended sequences, with recent works attempting to mitigate this by compressing context frames [21, 89].

Video world models. World models learn to predict state transitions resulting from actions, enabling application in agent planning and interactive simulations. Video prediction methods have been applied to learn visual world models. While early research [12, 23, 46, 48] typically relies on recurrent neural networks (RNNs) [30] and variational autoencoders [39], recent works have shifted towards more scalable transformer-based diffusion/AR models. Latest efforts in this direction have explored training such models without ground truth action data [8, 46], scaling to more complex worlds and controls [31, 66, 75, 80], enabling real-time inference [14, 16], and generating open-world environments [9, 16, 88]. To support continuous generation, all these approaches evaluate their Transformer models in an autoregressive, sliding-window fashion, which inherently limits the model's context to only a few recent frames, typically covering at most a few seconds. A direct consequence of this design is that these models inherently lack long-term coherence, a limitation explicitly pointed out by these works and directly experienced by practitioners [62].

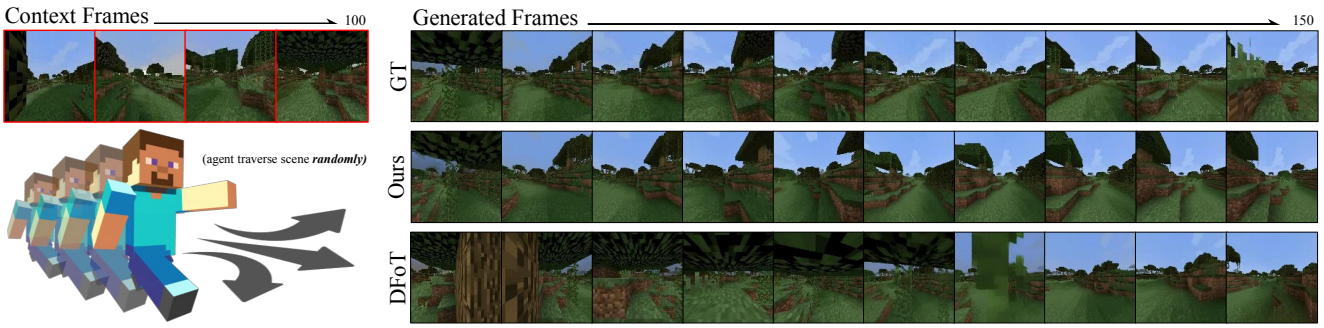

Figure 2. **Long memory video generation.** Our method generates sharp and consistent video predictions. Conditioned on agent actions, our method can accurately reconstruct previously visited regions of an environment, while maintaining linear training complexity and constant inference costs. In contrast, although state-of-the-art diffusion forcing transformers (DFoT [60]) can generate consistent looking videos over long horizons, their memory is bounded by the maximum context seen during training. Unlike our method which has linear scaling, attention-based transformers training scales quadratically, making it prohibitively expensive to train on longer videos.

Linear attention. Recently, linear RNNs have been proposed as an efficient alternative to self-attention [67]. The seminal work by Katharopoulos et al. [38] introduced linear attention that replaces softmax in attention with a kernel function, which reduces runtime complexity from quadratic to linear. Subsequent works have improved linear RNNs by refining the state update rule [4, 13, 20, 57, 82], enabling efficient parallelized [83] and hardware-aware training [13, 81], or incorporating more expressive hidden states [4, 64]. One notable family of linear RNNs is state-space models (SSMs) [13, 20, 49]. SSMs can be interpreted as a hybrid of convolutional neural networks and recurrent neural networks that combines the best of both worlds—parallelizable training and efficient autoregressive inference.

Recent efforts have introduced SSMs into image and video generation domains by replacing self-attention layers with efficient Mamba layers [19, 43, 44, 65, 69, 76, 77]. These approaches typically perform bidirectional scanning over the entire token sequence, therefore not utilizing the efficient autoregressive inference capabilities of SSMs. In contrast, our work employs unidirectional SSMs for modeling temporal dynamics and world state transitions, naturally leveraging their inherent advantagess.

3. Preliminaries

3.1. Video Diffusion Models

A diffusion process progressively corrupts an observed datapoint sampled from the data distribution $p(x_0)$ by adding Gaussian noise to it. The noisy data is given by:

$$x_t = \alpha_t x_0 + \sigma_t \epsilon, \tag{1}$$

where the scalar parameters $\alpha_t, \sigma_t > 0$ control the signal-to-noise ratio according to a predefined noise schedule and ϵ is sampled from a standard Gaussian distribution. Diffusion models [28, 59, 61] are typically trained to predict the noise by minimizing the following denoising objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left\| \epsilon_\theta(x_t, t) - \epsilon \right\|_2^2 \tag{2}$$

or alternative but equivalent targets such as the original clean data x_0 or the velocity $\epsilon - x_0$.

Diffusion models for video generation typically employ a two-stage approach: first encoding raw videos into latent space using a 3D variational autoencoder (VAE) [7, 22, 68], then learning a diffusion model in this latent space.

In conventional video diffusion models, the noise level is the same across all latent frames at each training iteration. This requires simultaneously generating all video frames during inference, where all frames following the same noise schedule. Diffusion forcing [10] introduces a strategy where noise levels are sampled independently per frame during training, enabling sequential video generation at inference time by denoising each frame conditioned on previously generated clean frames. The ability to generate a video autoregressively conditioned on streaming controls is essential for world modeling applications, such as gaming or robotic learning. Denote $\{x_0^i\}_{i=1}^T$ as a sequence of T latent frames. The noisy latent frames during training are obtained by

$$x_{t_i}^i = \alpha_{t_i} x_0^i + \sigma_{t_i} \epsilon^i, \tag{3}$$

where t_i is sampled independently for each frame i. Previous approaches have explored various backbone architectures under the diffusion forcing training scheme, including recurrent neural networks [10] and causal [87] or bidirectional [60] transformers.

The main component of the transformer architecture [67] is self-attention, where each token in a sequence attends to all others via dot-product similarity between learned query, key, and value representations. Given a sequence of input embeddings X, self-attention computes:

$$\text{Attn}(X) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V, \tag{4}$$

where Q, K, V are linear projections of X, and d is the latent dimension. Causal attention is a variant of self-attention that only allows tokens to attend to previous tokens

in the sequence. This is achieved by masking the attention matrix to prevent information flow from future tokens. In video diffusion models, previous approaches have applied block-wise causal mask to ensure that each token can only attend to tokens in the same or previous frames [87], thereby enabling autoregressive generation with KV-caching.

3.2. State-Space Models (SSMs)

An SSM models a sequence dynamic as a linear system

$$H_t = AH_{t-1} + BX_{t-1}; \quad X_t = CH_t + DX_{t-1}, \qquad (5)$$

where H_t are latent states, and A, B, C, D are matrices with appropriate dimensions. Building on this framework, Mamba [20] provides additional expressivity by modeling the parameters A, B, C, D as linear projections of input X_t, *i.e.,* $A_t := \text{Linear}_{\theta_A}(X_t)$ and similarly for B_t, C_t, and D_t. This allows the sequence dynamic to be content-aware, and can be viewed as a generalization of causal linear attention, where B and C play the role of K and Q, respectively.

Unlike attention, however, the generation of new tokens during inference takes the form of expression (5), and requires only the latent state H_T, where T is the most recent token. This offers superior time and space complexity compared to attention (see Tab. 1), but at the cost of compressed memory representation, as inference no longer involves all-to-all comparison of past inputs.

4. Methods

While SSMs have proven to be an effective drop-in replacements for attention in domains such as language modeling, directly substituting attention blocks with SSMs leads to suboptimal results for autoregressive video generation. In this section, we identify and address several shortcomings in the naive approach regarding ***architecture*** (Sec. 4.1), and ***training*** (Sec. 4.2). Our solutions result in a model that displays long-term spatial memory while maintaining ***constant per-frame inference cost*** throughout generation (Sec. 4.3).

4.1. Model Architecture

Because our model generates video frames autoregressively (one frame at a time), the temporal dimension (the sequence of frames) must be placed at the end of the scanning order. This "spatial-major/time-minor" ordering ensures that the model processes all spatial information within the current frame before moving to the next frame, thus preserving causal constraints and preventing the model from accessing future frame information. However, the spatial-major scan order makes it challenging to capture long-term temporal dependencies, as temporally adjacent tokens become distant from each other in the flattened token sequence. To address this limitation, we introduce a method that balances temporal memory and spatial coherence using a block-wise reordering of the spatio-temporal tokens.

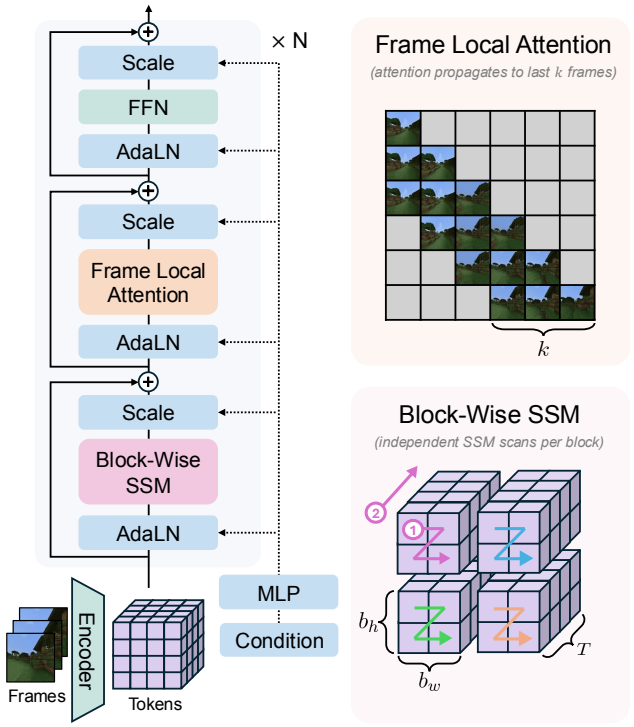

Figure 3. **Model architecture.** Our model features a block-wise SSM scan that divides spatial dimensions into independent scanning blocks (b_h, b_w), balancing temporal memory with spatial coherence. This works alongside frame local attention, which enables bidirectional processing within frames while maintaining causal relationships across the previous k frames, resulting in improved per-frame visual quality and temporal consistency.

Block-wise SSM scan. As shown in Fig. 3 (bottom right), our method breaks up the original sequence of tokens along the spatial dimensions into blocks of size (b_h, b_w, T), where b_h and b_w are layer-dependent block heights/width, and T is the temporal dimension of the data. Instead of performing a single scan over the entire sequence of tokens, a separate scan is performed for each token block. By controlling the values of b_h and b_w, we enable a trade-off between temporal correlation and spatial coherence. Temporally adjacent tokens are now separated by $b_h \times b_w$ tokens rather than $H \times W$ (as in conventional spatial-major scanning), where H, W represent the height/width of each frame. However, smaller blocks lead to worse spatial coherence, as the independent scans prevent tokens in different blocks from interacting. Therefore, the choice of block-size represents an effective way of trading off consistent long-term memory for short-term spatial consistency. Our model leverages the benefits of both small and large block sizes, by employing different values for b_h and b_w in different layers.

SSMs can struggle with high-complexity tasks such as visual generation due to the limited expressivity of the fixed-dimensional SSM state. Our block-wise scanning method mitigates this limitation by effectively increasing

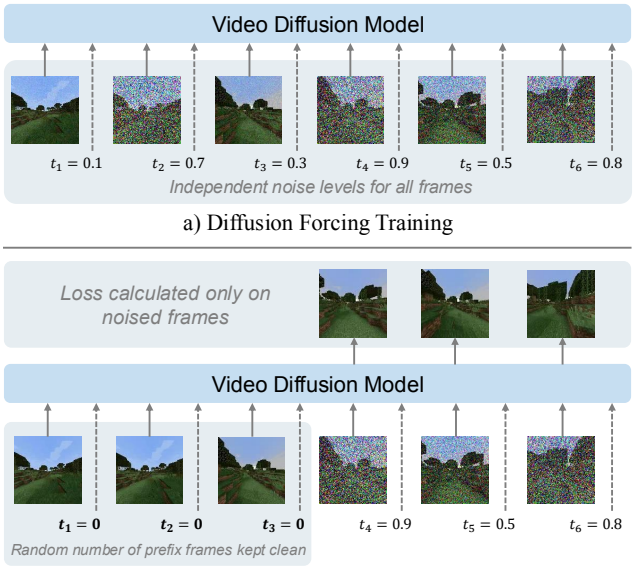

Figure 4. **Improved long-context training.** a) Standard diffusion forcing injects independent noise levels to all frames, b) Our method keeps a random number of initial frames completely clean ($t_i = 0$), adds independent noise to later frames, and calculates loss only on the noised frames. By providing clean context frames distant from denoising targets, our approach encourages the model to learn long-term dependencies.

the dimensionality of the SSM state at each layer, as each block is allocated a separate state.

Frame local attention. Linear attention variants such as Mamba have shown to struggle in tasks related to associative recall [82]. In video generation, the inability of Mamba to retrieve precise local information results in poor frame-wise quality and loss of short-term temporal consistency. Prior works [4, 55, 82, 83] have shown hybrid architectures that combine local attention with SSMs can improve language modeling. In our model, we introduce a frame-wise local attention block following every Mamba scan, as illustrated in Fig. 3. During training, we apply block-wise causal attention where each token can only attend to tokens in the same frame and a fixed-size window of previous frames. The attention mask M takes the form,

$$M_{i,j} = \begin{cases} 1, & \text{if } j \in [i-k, i], \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

where i and j are indices to frames in the sequence, and k is the window size.

Action condition. We enable interactive controls during autoregressive generation by passing actions corresponding to each frame as input. Continuous action values (e.g., camera positions) are processed through a small MLP and added to the noise level embeddings, which are then injected into the network via adaptive normalization layers [3, 33, 34, 52]. For discrete actions, we directly learn embeddings corresponding to each possible action.

4.2. Long-Context Training

Although our architecture design enhances the model's ability to maintain long-term memory, it is still challenging to learn long temporal dependencies with standard diffusion training schemes. Video data contains significant redundancy, allowing models to rely primarily on nearby frames for denoising in most cases. As a result, diffusion models frequently become trapped in local minima, failing to capture long-term dependencies.

Standard diffusion forcing always adds noise independently to each frame during training. Under these conditions, the model has limited incentive to reference distant context frames since they usually contain less useful information than local frames. To encourage the model to attend to distant frames and learn long-term correlations, we mix diffusion forcing with a modified training scheme that maintains a random-length prefix of frames completely clean (noise-free) during training, as illustrated in Fig. 4. When large noise is added to the later frames, the clean context frames may provide more useful information than the noisy local frames, prompting the model to utilize them effectively. This is similar to the training scheme in Ca2-VDM [18] although our motivation is different and we still keep independent noise level for the later noisy frames.

4.3. Efficient Inference via Fixed-Length State

During inference, we autoregressively generate new video frames conditioned on input actions. Our hybrid architecture ensures constant speed and memory usage. Specifically, each layer of our model only tracks: (1) a fixed-length KV-cache for the previous k frames, and (2) the SSM state for each block. This ensures constant memory usage throughout the generation process, unlike fully causal transformers whose memory requirements grow linearly as they store KV-caches for all previous frames. Similarly, our method maintains constant per-frame generation speed, as local attention and block-wise SSM computations do not scale with video length. This property is crucial for video world model applications, where generating video frames indefinitely without performance degradation is essential.

5. Experiments

We evaluate our method in terms of training and inference efficiency, as well as long-term memory capabilities. To this end, we utilize two long video datasets and evaluate the model's performance on spatial memory tasks that require recalling information from distant frames to generate accurate predictions. Details on the datasets are provided in Sec. 5.1, metrics in Sec. 5.2 and results in Sec. 5.3.

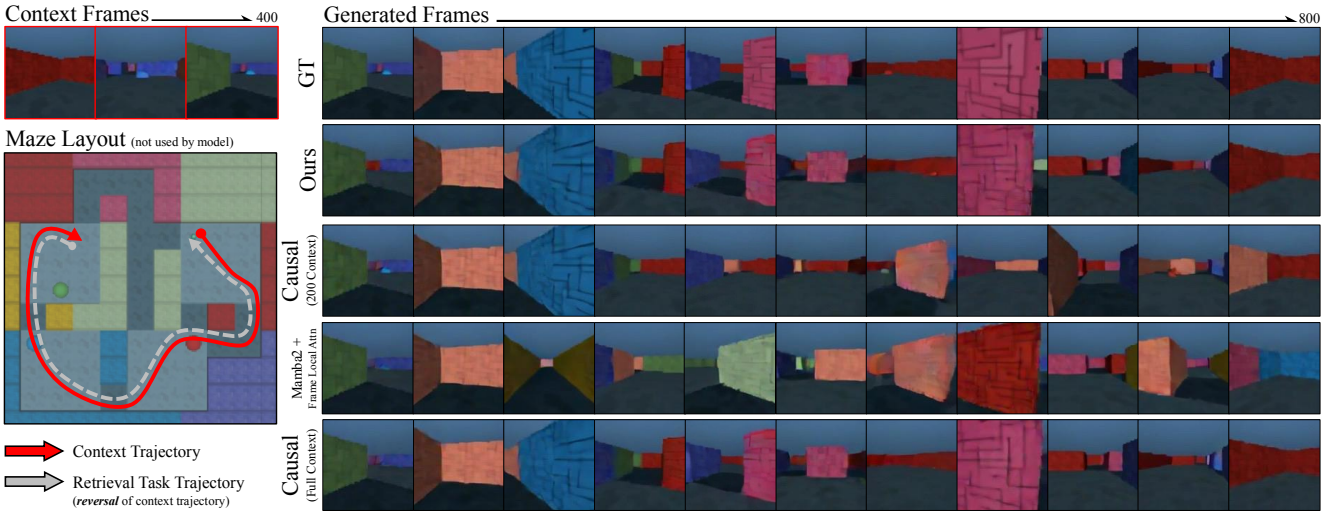

Figure 5. **Overview of retrieval task and qualitative results.** The maze layout shows the context trajectory (red) and retrieval trajectory (gray), which reverses the original path. We compare each model on a retrieval task with 400 generated frames following 400 context frames. Top row displays ground truth frames. Our model demonstrates high fidelity to ground truth throughout the sequence. The Causal model with 192-frame context deteriorates quickly beyond its training window, while Mamba2 + frame local attention fails completely. The Causal model with full context performs well but requires quadratic complexity during training and linear complexity during inference.

5.1. Datasets

Memory Maze. Memory Maze [51] is a 3D domain of randomized mazes, designed for evaluating long-term memory capabilities of RL agents. Samples from the dataset contains trajectories collected from an agent exploring a randomly generated maze. Each trajectory contains 2000 action-frame pairs, which specify the most recent action/position of the agent, along with the observation of the maze from the agent's point-of-view. We train all models on position/observation pairs. Any information regarding the ground truth layout of the maze is excluded.

TECO Minecraft. The TECO [79] dataset consists of 200K gameplay trajectories collected from Minecraft. During data collection, an agent takes one of four actions (forward, turn left, turn right, jump) in random sequences, producing a trajectory consisting 150[1] action/observation pairs. These random action sequences result in trajectories where the agent occasionally revisit regions seen earlier.

5.2. Evaluations

Spatial retrieval task. This task involves providing the model with a random agent trajectory and the corresponding observations as context, then tasking the model to backtrack through the exact sequence of actions to the agent's starting position. Given the scene is static, the generated sequence should reverse the context frames. We refer to this as a retrieval task because the ground truth answer can be retrieved from the given context frames. Fig. 5 shows an example of the context and generation trajectories. We evaluate the retrieval task only on the Maze dataset, as certain actions in Minecraft are not invertible (e.g. jump).

[1]The original dataset contains 300 action/observation pairs, following [60], we sample every other frame.

Model	Retrieval (400 Frames)		
	SSIM ↑	LPIPS ↓	PSNR ↑
Causal (192 Frame Context)	0.829	0.147	26.4
Mamba2 [13]	0.747	0.313	20.4
Mamba2 + Frame Local Attn	0.735	0.336	19.3
Ours	**0.898**	**0.069**	**30.8**
Causal (Full Context)	0.914	0.057	32.6

Table 2. **Quantitative results on retrieval task.** Comparison of model performance on the 400-frame retrieval task using SSIM, LPIPS, and PSNR. Our model outperforms all baselines with sub-quadratic complexity, while approaching the performance of full-context causal transformers with quadratic training complexity.

Spatial reasoning task. Similar to the retrieval task, the spatial reasoning task involves providing the model with a random agent trajectory and observations as context. However, instead of backtracking, the model continues the trajectory with random actions. Assuming the model has been given enough context such that the entire environment has been committed into memory, the model should reconstruct every observation along the continued trajectory. Fig. 6 shows an example of the context and generated trajectories.

5.3. Results

Results on Memory Maze. Tables 2 and 3 present quantitative comparisons in terms of spatial retrieval and reasoning on Memory Maze. We evaluate each model by comparing generated frames against ground truth using similarity metrics including SSIM [72], LPIPS [90], and PSNR [35].

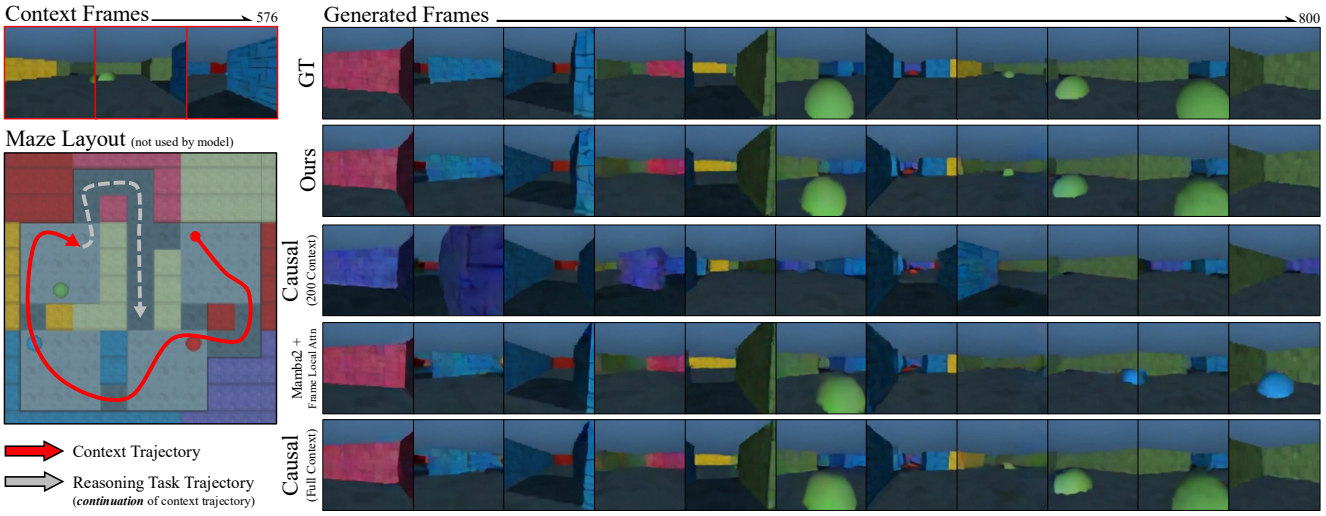

Figure 6. **Overview of reasoning task and qualitative results.** The maze layout illustrates the context trajectory (red) and reasoning trajectory (gray)[1], which continues the context path. We compare each model on a retrieval task with 224 generated frames conditioned on 576 context frames. The causal transformer trained on 192 frames fails immediately, as the queried region lies beyond its trained context. Mamba2 + Frame Local Attention fails to recall visual details such as positions of the balls. Our method successfully reconstructs previously visited regions of the maze, with performance comparable to the full-context causal transformer.

Model	Reasoning (224 Frames)		
	SSIM ↑	LPIPS ↓	PSNR ↑
Causal (192 Frame Context)	0.839	0.125	27.1
Mamba2 [13]	0.827	0.150	26.4
Mamba2 + Frame Local Attn	0.845	0.113	27.5
Ours	**0.855**	**0.099**	**28.2**
Causal (Full Context)	0.860	0.089	28.8

Table 3. **Quantitative results on reasoning task.** Performance comparison on the 224-frame reasoning task conditioned on 576 context frames. Our model surpasses all other sub-quadratic methods and performs close to full-context causal transformers.

For both tasks, we compare our method against baselines with sub-quadratic training complexity. These include a causal transformer trained on a limited context length[2], an architecture with only Mamba2 blocks, and Mamba2 + Frame Local Attn, a hybrid model that combines our frame local attention with Mamba2. We also include results from a causal transformer trained on the full context as reference. Our model outperforms all sub-quadratic baselines across all metrics in both tasks. While the causal transformer with full context gives the best performance, it comes with quadratic training and inference complexity. As shown in Figures 5 and 6, frame predictions from other sub-quadratic models deviate from the ground truth after a certain period for both tasks, whereas our method maintains accurate predictions throughout the trajectory.

In Fig. 7, we further analyze the performance of each method on the retrieval task, showing the change of retrieval accuracy as the distance between generated and retrieved frames increases. Causal transformers perform well within their training context but drop off quickly beyond their maximum training length. Other linear-complexity methods such as Mamba and Mamba2 + Frame Local Attn perform poorly due to limited state-space expressivity. In contrast, our method maintains high accuracy across all retrieval distances, comparable to a causal transformer trained on the full context.

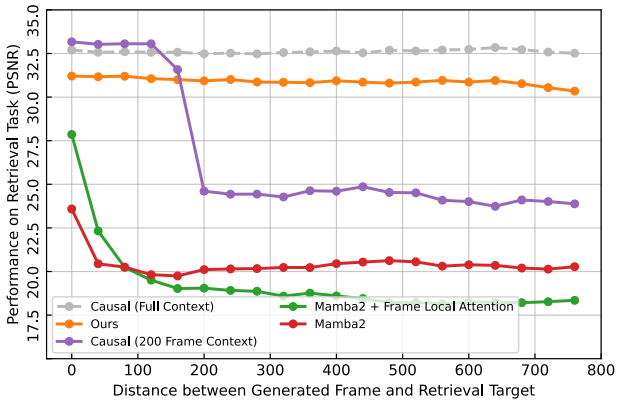

Figure 7. **Retrieval PSNR vs. Frame Distance.** Our model maintains consistent high performance comparable to full-context transformers while significantly outperforming limited-context transformers that degrade beyond training length and linear-complexity models that lack sufficient expressivity throughout.

Results on TECO Minecraft. We show quantitative and qualitative results on the Minecraft dataset in Tab. 4 and Fig. 2 respectively. We compare our method to diffu-

[1]Trajectories shown in the figure are for illustrative purposes only. Actual trajectories tend to be longer, covers the whole maze, and revisits regions multiple times.

[2]Since the causal transformer has never seen the full context during training, we consider it sub-quadratic.

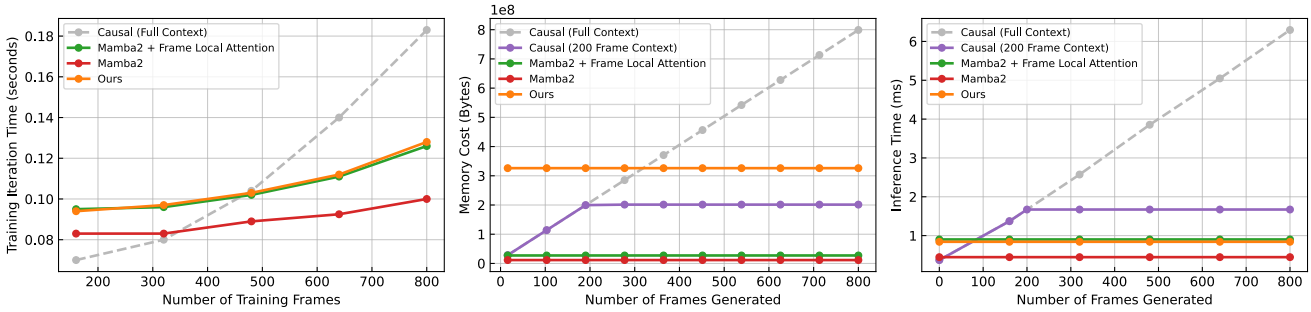

Figure 8. **Training and inference performance comparisons.** Evaluation of training costs (left), inference memory usage (center), and inference time (right), demonstrating how our approach maintains consistent memory and computational efficiency as frame count increases compared to baseline methods.

Model	Reasoning (50 Frames)		
	SSIM ↑	LPIPS ↓	PSNR ↑
DFoT [60]	0.450	0.281	17.1
Causal (25 Frame Context)	0.417	0.350	15.8
Ours	**0.454**	**0.259**	**17.8**

Table 4. **Reasoning task results on Minecraft dataset.** Performance comparison on the 50-frame reasoning task conditioned on 100 frames. Across all metrics, our model outperforms both DFoT [60] and causal transformers limited to 25-frame context.

Model	Reasoning (200 Frames)		
	SSIM ↑	LPIPS ↓	PSNR ↑
Ours w/o block-wise scan	0.845	0.113	27.5
Ours with block size 1	0.766	0.198	23.1
Ours w/o Sec. 4.2	0.809	0.143	25.3
Ours (Full)	**0.855**	**0.099**	**28.2**

Table 5. **Ablations.** Performance of various setups of our model on the maze reasoning task. Every component, from architecture to training, is crucial for achieving accurate long-context memory.

sion forcing transformer [60] (DFoT), a bidirectional transformer trained under the diffusion forcing regime. DFoT represents the current state-of-the-art architecture in autoregressive long video generation. However, due to the quadratic complexity of their model, DFoT is trained on a limited context length of 25 frames. As shown in Fig. 2, our method can accurately predict previously explored regions, as opposed to methods with a limited context window.

Our method outperforms both DFoT and a causal transformer trained on a 25-frame context. All models obtain lower similarity on this dataset due to shorter trajectories, where the model is given only 100 frames of context to predict 50 frames. Often, a 100-frame context is insufficient for the agent to fully observe the environment, causing task trajectories to venture into previously unseen regions in which case frame-wise similarity becomes less informative.

Training and inference costs. Fig. 8 evaluates model performance using three metrics: training cost per iteration (left), memory utilization during generation (center), and computational time during inference (right). Our method demonstrates superior scaling across all metrics, with training time scaling linearly with context length while maintaining constant memory and computational costs during inference. For inference runtime comparison, we compare the runtime of a single forward pass through our frame local attention plus SSM update against full attention with KV-caching across all previously generated frames.

5.4. Ablations.

Block-wise SSM scan. Tab. 5 shows the benefits of our block-wise SSM scan. Without it and with only a single scan along all spatiotemporal tokens, the model struggles with maintaining memory over long horizons due to the lack of proximity in adjacent temporal tokens and limited SSM state capacity. Conversely, always using the smallest possible block size (i.e., $b_h, b_w = 1$) ensures temporal token proximity but sacrifices spatial correlations, resulting in poor performance on the reasoning task where temporal retrieval alone is insufficient.

Long-context training. Tab. 5 highlights the importance the training scheme outlined in Sec. 4.2. Without this adjustment to diffusion forcing, our model falls into a local minimum and generates frames without looking at distant context, resulting in poor spatial reasoning performance.

6. Limitations and Future Work

While our work represents a significant step towards scalable and consistent world models, it has certain limitations. First, despite achieving constant inference time, our method does not yet support interactive frame rates. Future work could speed-up the generation through timestep distillation [87]. Second, our method cannot effectively handle memory longer than the training context length. Recent works [5, 86] on length extrapolation for Mamba architectures could potentially extend our memory capacity. Finally, our experiments are limited to low-resolution synthetic videos due to computational constraints. Scaling up to high-resolution, realistic videos is left for future work.

References

[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *NeurIPS*, 2024. 1

[2] Genesis Authors. Genesis: A universal and generative physics engine for robotics and beyond, 2024. 1

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*, 2024. 3, 5

[5] Assaf Ben-Kish, Itamar Zimerman, Shady Abu-Hussein, Nadav Cohen, Amir Globerson, Lior Wolf, and Raja Giryes. Decimamba: Exploring the length extrapolation potential of mamba. 2025. 8

[6] 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. 2

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 3

[8] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie Chan, Nicolas Manfred Otto Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. Genie: Generative interactive environments. *ArXiv*, abs/2402.15391, 2024. 2

[9] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *ICLR*, 2025. 1, 2

[10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024. 2, 3

[11] 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. In *ICLR*, 2023. 2

[12] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. In *ICLR*, 2017. 2

[13] Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024. 3, 6, 7

[14] Julian Decart, Quinn Quevedo, Spruce McIntyre, Xinlei Campbell, Robert Chen, and Wachen. Oasis: A universe in a transformer. 2024. 1, 2

[15] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels. *ArXiv*, abs/2412.05496, 2024. 14

[16] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024. 1, 2

[17] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, and Jun Xiao. Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models. *arXiv preprint arXiv:2406.10981*, 2024. 2

[18] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *arXiv preprint arXiv:2411.16375*, 2024. 1, 2, 5

[19] Yu Gao, Jiancheng Huang, Xiaopeng Sun, Zequn Jie, Yujie Zhong, and Lin Ma. Matten: Video generation with mamba-attention. *arXiv preprint arXiv:2405.03025*, 2024. 3

[20] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *COLM*, 2024. 1, 3, 4

[21] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *ArXiv*, abs/2503.19325, 2025. 2

[22] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Fei-Fei Li, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. In *ECCV*, 2024. 3

[23] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018. 1, 2

[24] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2

[25] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020. 1

[26] Haoran He, Yang Zhang, Liang Lin, Zhongwen Xu, and Ling Pan. Pre-trained video generative models as world simulators. *arXiv preprint arXiv:2502.07825*, 2025. 1

[27] 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*, 2022. 2

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022. 1, 2

[30] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[31] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023. 1, 2

[32] Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. *arXiv preprint arXiv:2412.07720*, 2024. 1, 2

[33] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 5

[34] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018. 5

[35] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 2008. 6

[36] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025. 1, 2

[37] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 1

[38] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *ICML*, 2020. 3

[39] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 2

[40] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In *ICML*, 2024. 2

[41] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 2

[42] Muyang Li, Ji Lin, Chenlin Meng, Stefano Ermon, Song Han, and Jun-Yan Zhu. Efficient spatially sparse inference for conditional gans and diffusion models. *Advances in neural information processing systems*, 35:28858–28873, 2022. 2

[43] Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Clear: Conv-like linearization revs pre-trained diffusion transformers up. *arXiv preprint arXiv:2412.16112*, 2024. 3

[44] Songhua Liu, Weihao Yu, Zhenxiong Tan, and Xinchao Wang. Linfusion: 1 gpu, 1 minute, 16k image. *arXiv preprint arXiv:2409.02097*, 2024. 3

[45] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025. 1

[46] Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *CVPR*, 2021. 2

[47] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *ICLR*, 2023. 1

[48] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *NeurIPS*, 2015. 2

[49] Yuta Oshima, Shohei Taniguchi, Masahiro Suzuki, and Yutaka Matsuo. Ssm meets video diffusion models: Efficient long-term video generation with structured state spaces. *arXiv preprint arXiv:2403.07711*, 2024. 2, 3

[50] Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. 2024. 1, 2

[51] Jurgis Pasukonis, Timothy Lillicrap, and Danijar Hafner. Evaluating long-term memory in 3d mazes. *arXiv preprint arXiv:2210.13383*, 2022. 6

[52] William S Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 5

[53] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1

[54] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[55] Liliang Ren, Yang Liu, Yadong Lu, Yelong Shen, Chen Liang, and Weizhu Chen. Samba: Simple hybrid state space models for efficient unlimited context language modeling. *ArXiv*, abs/2406.07522, 2024. 5

[56] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[57] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *ICML*, 2021. 3

[58] 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. In *ICLR*, 2023. 2

[59] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 3

[60] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025. 1, 2, 3, 6, 8, 13

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based

generative modeling through stochastic differential equations. In *ICLR*, 2021. 3

[62] stealthispost. Me playing a few minutes of ai minecraft. 2025. 2

[63] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024. 2

[64] Yu Sun, Xinhao Li, Karan Dalal, Jiarui Xu, Arjun Vikram, Genghan Zhang, Yann Dubois, Xinlei Chen, Xiaolong Wang, Sanmi Koyejo, et al. Learning to (learn at test time): Rnns with expressive hidden states. *arXiv preprint arXiv:2407.04620*, 2024. 3

[65] Yao Teng, Yue Wu, Han Shi, Xuefei Ning, Guohao Dai, Yu Wang, Zhenguo Li, and Xihui Liu. Dim: Diffusion mamba for efficient high-resolution image synthesis. *arXiv preprint arXiv:2405.14224*, 2024. 3

[66] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 1, 2

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[68] R Villegas, H Moraldo, S Castro, M Babaeizadeh, H Zhang, J Kunze, PJ Kindermans, MT Saffar, and D Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *ICLR*, 2023. 3

[69] Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, et al. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. *arXiv preprint arXiv:2412.09856*, 2024. 3

[70] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024. 1

[71] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024. 2

[72] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 6

[73] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *CVPR*, 2024. 1

[74] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *CoRL*, 2023. 1

[75] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 2

[76] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 3

[77] Jing Nathan Yan, Jiatao Gu, and Alexander M Rush. Diffusion models without attention. In *CVPR*, 2024. 3

[78] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 2

[79] Wilson Yan, Danijar Hafner, Stephen James, and P. Abbeel. Temporally consistent transformers for video generation. In *ICML*, 2022. 6

[80] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023. 2

[81] Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. *arXiv preprint arXiv:2312.06635*, 2023. 3

[82] Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving mamba2 with delta rule. *arXiv preprint arXiv:2412.06464*, 2024. 3, 5

[83] Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *Advances in Neural Information Processing Systems*, 37:115491–115522, 2025. 3, 5

[84] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 13

[85] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 1

[86] Zhifan Ye, Kejing Xia, Yonggan Fu, Xin Dong, Jihoon Hong, Xiangchi Yuan, Shizhe Diao, Jan Kautz, Pavlo Molchanov, and Yingyan Celine Lin. Longmamba: Enhancing mamba's long-context capabilities via training-free receptive field enlargement. In *ICLR*, 2025. 8

[87] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025. 1, 2, 3, 4, 8

[88] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025. 1, 2

[89] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *ArXiv*, abs/2504.12626, 2025. 2

[90] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[91] Zhicheng Zhang, Junyao Hu, Wentao Cheng, Danda Paudel, and Jufeng Yang. Extdm: Distribution extrapolation diffusion model for video prediction. In *CVPR*, 2024. 2

[92] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. Is sora a world simulator? a comprehensive survey on general world models and beyond. *arXiv preprint arXiv:2405.03520*, 2024. 1