Context as Memory: Scene-Consistent Interactive Long Video Generation with Memory Retrieval

Jiwen Yu[1†] Jianhong Bai[2†] Yiran Qin[1]
Quande Liu[3‡] Xintao Wang[3] Pengfei Wan[3] Di Zhang[3] Xihui Liu[1‡]
[1] The University of Hong Kong [2] Zhejiang University [3] Kuaishou Technology
https://context-as-memory.github.io/

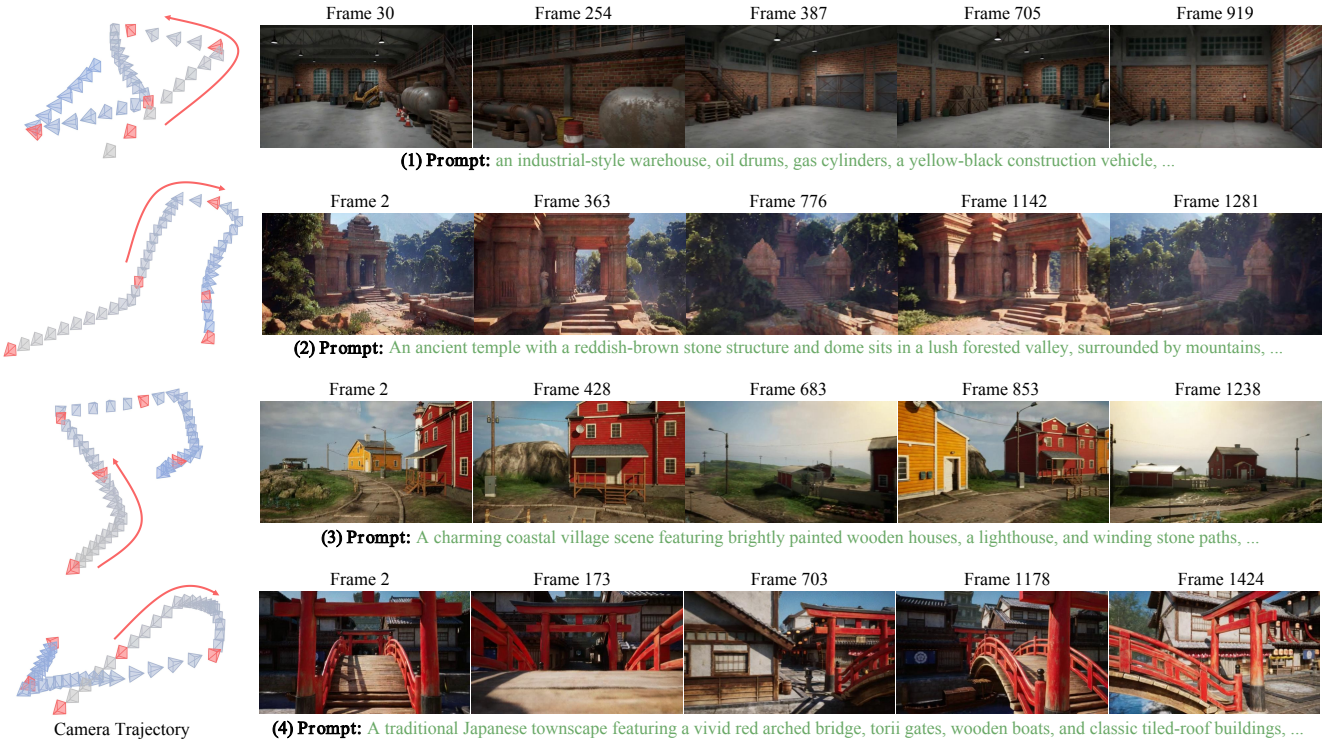


Figure 1. We propose **Context-as-Memory**, which leverages historical frames to ensure scene consistency in long video generation by guiding the synthesis of new frames. Key frames from generated videos along camera trajectories are shown, with camera poses marked in red. It can be observed that different frames maintain good consistency when viewing the same scene from different viewpoints.

Abstract

Recent advances in interactive video generation have shown promising results, yet existing approaches struggle with scene-consistent memory capabilities in long video generation due to limited use of historical context. In this work, we propose ***Context-as-Memory****, which utilizes historical context as memory for video generation. It includes two simple yet effective designs: (1) storing context in frame format without additional post-processing; (2) conditioning by concatenating context and frames to be predicted along the frame dimension at the input, requiring no external control modules. Furthermore, considering the enormous computational overhead of incorporating all historical context, we propose the* ***Memory Retrieval*** *module to select truly relevant context frames by determining FOV (Field of View) overlap between camera poses, which significantly reduces the number of candidate frames without substantial information loss. Experiments demonstrate that* ***Context-as-Memory*** *achieves superior memory capabili-*

‡Corresponding authors. †Work done during an internship at KwaiVGI, Kuaishou Technology.

ties in interactive long video generation compared to SOTAs, even generalizing effectively to open-domain scenarios not seen during training. The link of our project page is https://context-as-memory.github.io/.

1. Introduction

Recent breakthroughs in video generation models [22, 28, 32, 39, 47] have shown remarkable progress. Due to their powerful generative capabilities developed through training on large-scale real-world datasets, these models are considered to have the potential to become world models capable of modeling reality [28, 30, 45, 46]. Among various research directions in this field, interactive long video generation has emerged as a crucial one since many applications, such as gaming [6, 38, 52] and simulation [11, 17, 33], require interactive long video generation, where the videos are generated in a streaming manner controlled by user interactions. Recent works on long video generation [4, 8, 12, 21, 34, 40, 54] have facilitated research in this field.

Despite these advances, current approaches still face significant challenges in terms of memory capabilities [50, 51], which refers to a model's ability to maintain content consistency during continuous video generation, such as preserving the scene when the camera returns to a previously viewed location. Take Oasis [5] as an example: while it can generate lengthy Minecraft gameplay videos, even simple operations like turning left and then immediately right result in completely different scenes. This issue is prevalent across various state-of-the-art methods [18, 34, 38, 52], suggesting that while current approaches can generate videos of extended duration, they struggle to maintain coherent long-term memory of scene content and spatial relationships.

In our view, these methods' limitations in memory capabilities are not surprising. This is because when generating each new video frame, these methods can only predict based on a limited number of previous frames. For instance, Diffusion Forcing [4, 34] can only utilize context from a fixed window of several dozen frames. While this setup works for video continuation, it fails to maintain long-term consistency. In the case of video generation, if each frame to be generated could reference all previously generated frames, the generative model could select and replicate relevant content from historical frames into the current frame being generated, thus it would be possible to maintain scene consistency in long videos. In other words, **all previously generated context frames serve as the memory.**

However, the idea of "all historical context as memory" seems intuitive but is impractical for three main reasons: (1) Including all historical frames in computation would be extremely resource-intensive. (2) Processing all historical frames is computationally wasteful since only a small fraction is relevant to the current frame generation. (3) Processing irrelevant historical frames adds noise that may hinder rather than help current frame generation. Therefore, a reasonable approach is to retrieve a small number of relevant frames from historical context as conditions for current generation, which we call "**Memory Retrieval**".

In this work, we propose **Context-as-Memory** as a solution for scene-consistent interactive long video generation, which includes two simple yet effective designs: (1) Storage format: directly store generated context frames as memory, requiring no post-processing such as feature embedding extraction or 3D reconstruction; (2) Conditioning method: directly incorporate as part of the input through concatenation for context learning, without requiring additional control modules like external adapters or cross attention. To effectively reduce unnecessary computational overhead and only condition on truly relevant context, we propose **Memory Retrieval**. Specifically, we introduce a rule-based approach based on camera trajectories. With a camera-controlled video generation model, we can annotate all context frames with camera information based on user's camera control. We can determine co-visibility by checking the FOV (Field of View) overlap based on camera poses at each timestamp along the trajectory, and then use this co-visibility relationship to decide which relevant frames to retrieve. To implement this solution, we collected a new scene-consistent memory learning dataset using Unreal Engine 5, featuring long videos with precise camera annotations across diverse scenes and camera trajectories. The same regions are captured across different viewpoints and times, enabling both FOV-based retrieval and long-term consistency supervision.

Our main contributions can be summarized as follows:

- We propose **Context-as-Memory**, highlighting the storage of frames as memory and conditioning via historical context learning for scene-consistent video generation.
- To utilize relevant history frames efficiently, we design **Memory Retrieval**, a rule-based approach using FOV overlap of camera trajectory.
- We introduce a long, scene-consistent video dataset with precise camera annotations for memory training, featuring diverse scenes and captions.
- Our experiments show superior long video generation memory, significantly outperforming SOTAs and achieving memory even in unseen, open-domain scenarios.

2. Related Work

2.1. Interactive Long Video Generation

Video Generation Model. Video generation models can generate video sequences $\mathbf{x} = \{x^0, x^1, \ldots, x^t\}$, where x^i indicates the i-th frame. The current mainstream architecture is based on diffusion models [16, 24, 25, 35, 36], which excel in generating high-quality content and have been

widely adopted [3, 7, 20, 22, 28, 32, 39, 47]. Other alternative architectures include next-token prediction [21, 40, 44] and various hybrid approaches [4, 8, 23].

Controllable Video Generation. This task can be formulated as $p(\mathbf{x}|c)$, where c represents different types of control signals. The most representative control signals include: camera motion control [1, 2, 10, 14, 41], and agent action control in games or simulators [5, 6, 9, 38, 52]. These control signals greatly enhance user interactive experience, enabling free exploration in the created virtual worlds.

Streaming Video Generation. Streaming video generation can condition on previously generated frames to continuously generate new video frames, which can be expressed as $p(x^0, x^1, ..., x^n) = \prod_{i=0}^{n} p(x^i|x^0, x^1, ..., x^{i-1})$, where x^i indicates the i-th frame. Representative approaches include Diffusion-based methods [4, 12, 34, 52, 54] and GPT-like next token prediction methods [18, 21, 40]. Diffusion-based methods generally achieve higher visual quality and faster sampling speed, thus we focus on diffusion models for long video generation in this work. Although these SOTA methods generally fail to generate long videos with scene-consistent memory, instead only producing long videos with short-term continuity.

Memory Capability for Video Generation. Many related works' demos [5, 18, 34, 38] have shown that current long video generation methods generally lack memory capability: while maintaining frame-to-frame continuity, the scenes continuously change. One potential approach [26, 31, 49, 53] is to leverage 3D reconstruction to build explicit 3D representations from generated videos, then render initial frames from these 3D representations as conditions for new video generation. However, this method is limited by the accuracy and speed of 3D reconstruction, particularly in continuously expanding large scenes where accumulated 3D reconstruction errors become intolerable. Moreover, these works focus on 3D generation and merely borrow priors from video generation models, which differs from our scope. WorldMem [42] attempts to implement memory by injecting historical frames through cross attention, and has been validated on video lengths of around 10 seconds in Minecraft scenarios.

2.2. Context Learning for Video Generation

Recently, some works [12, 13, 54] have begun to explore the role of long-context in video generation. LCT [13] performs long-context tuning on pre-trained single-shot video diffusion models to achieve consistency in multi-shot video generation. FAR [12] proposes Long-Term and Short-Term context windows to condition video generation models for long video generation. FramePack [54] introduces a hierarchical method to compress context frames into a fixed number of frames as conditioning for video generation models to achieve long video generation. However, their compression method loses too much information from temporally distant frames. In this work, we further highlight the significance of context, emphasizing that all history context serves as memory for scene-consistent long video generation.

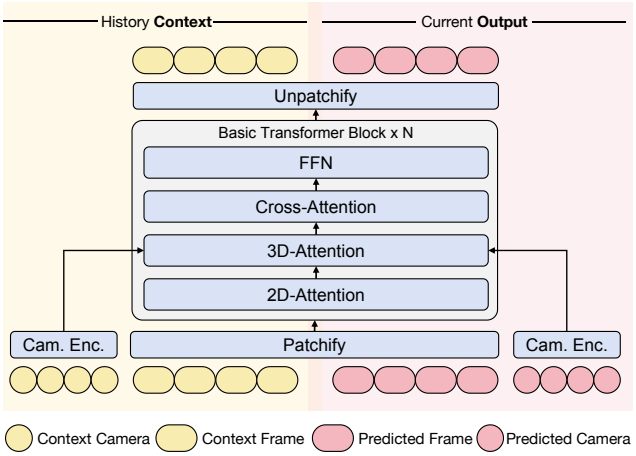


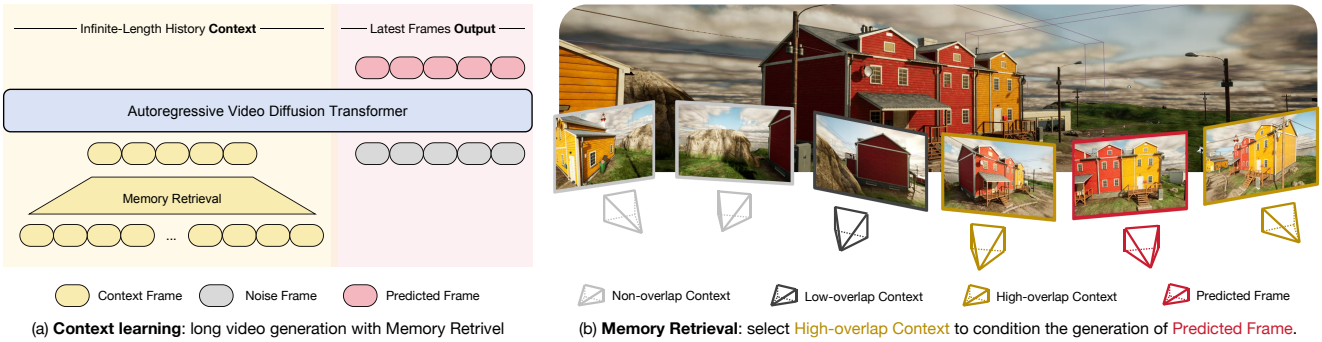
Figure 2. **Model Architecture.** We concatenate the context frames to be conditioned and the predicted frames along the frame dimension. This method of injecting context is simple and effective, requiring no additional modules.

3. Method

As discussed in Section 1, we propose that historical context frames can serve as memory for scene-consistent interactive long video generation. This section will detail how we implement this approach. Specifically: Sec. 3.1 introduces preliminaries. Sec. 3.2 describes how to inject context frames as conditions for video generation. Sec. 3.3 presents our Memory Retrieval method, which selects most relevant context frames to guide the generation of new frames. This section includes alternative approaches and our proposed search method based on camera trajectories. Sec. 3.4 introduces our long video dataset collected using Unreal Engine 5, which features precise camera pose annotations, diverse scenes, and caption annotations.

3.1. Preliminaries

Full-Sequence Text-to-Video Base Model. Our work is based on a full-sequence text-to-video model, specifically, a latent video diffusion model consisting of a causal 3D VAE [19] and a Diffusion Transformer (DiT) [29]. Each DiT block sequentially consists of spatial (2D) attention, spatial-temporal (3D) attention, cross-attention, and FFN modules. Let $\mathbf{x}$ represent a sequence of video frames, the Encoder of 3D VAE compresses it temporally and spatially to obtain the latent representation $\mathbf{z} = Encoder(\mathbf{x})$. With a temporal compression factor of r, the original $1+nr$ frames of $\mathbf{x} = \{x^0, x^1, ...x^{nr}\}$ are compressed into $1 + n$ latents



(a) **Context learning**: long video generation with Memory Retrivel

(b) **Memory Retrieval**: select High-overlap Context to condition the generation of Predicted Frame.

Figure 3. **Method Demonstration.** (a) We propose **Context-as-Memory**, where all historical context frames serve as memory conditions in the generation of predicted frames, with Memory Retrieval extracting relevant information from all context frames. (b) Our proposed **Memory Retrieval** method is a search algorithm based on camera trajectories. It selects relevant frames by evaluating the overlap between camera views of different frames.

of $\mathbf{z} = \{z^0, z^1, ..., z^n\}$. During training, random Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is added to the clean latent $\mathbf{z}_0$ to obtain noisy latent $\mathbf{z}_t$ at timestep t. The network $\epsilon_\phi(\cdot)$ is trained to predict the added noise, with the following loss function:

$$\mathcal{L}(\phi) = \mathbb{E}[||\boldsymbol{\epsilon}_\phi(\mathbf{z}_t, \mathbf{p}, t) - \boldsymbol{\epsilon}||], \tag{1}$$

where ϕ represents the parameters and $\mathbf{p}$ is the given text prompt. Then we use the predicted noise $\boldsymbol{\epsilon}_\phi$ to denoise the noisy latent. During inference, a clean latent $\mathbf{z}$ is sampled from a random Gaussian noise, then the Decoder of 3D VAE decodes it into video sequence $\mathbf{x} = Decoder(\mathbf{z})$.

Camera-Conditioned Video Generation. In our work, we incorporate camera control mechanisms [2, 41] into the video generation model to implement interactive video generation. By providing camera trajectories as conditioning for video generation, we can know the camera poses of each context frame in advance. Let $\mathbf{cam}$ represent the camera poses, where f denotes the total number of frames. Following the mechanism proposed in ReCamMaster [2], in order to inject $\mathbf{cam} = [R, t] \in \mathbb{R}^{f\times(3\times4)}$, we first map it to the same dimension as the model's feature channels through a camera encoder $\mathcal{E}_c(\cdot)$, followed by adding them together:

$$\mathbf{F}_i = \mathbf{F}_o + \mathcal{E}_c(\mathbf{cam}), \tag{2}$$

where $\mathbf{F}_o$ is the output of spatial attention, $\mathbf{F}_i$ is the input of 3D attention and $\mathcal{E}_c(\cdot)$ is one layer of MLP with ϕ_{MLP} as learnable parameters. During the training of camera control, we use the original diffusion loss as follows:

$$\mathcal{L}_{\mathbf{cam}}(\phi, \phi_{MLP}) = \mathbb{E}[||\boldsymbol{\epsilon}_{\phi,\phi_{MLP}}(\mathbf{z}_t, \mathbf{p}, \mathbf{cam}, t) - \boldsymbol{\epsilon}||]. \tag{3}$$

3.2. Context Learning Mechanism for Memory

Suppose the latent of context to be conditioned is $\mathbf{z}^c$, and we need to learn the conditional denoiser $p(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}^c)$. Considering that the context grows continuously during the generation process (i.e., the context is variable-length), methods designed for single-frame or fixed-length frame conditions, such as Adapter [27, 55] and channel-wise concatenation [43], are not applicable. Similar to ReCamMaster [2], we propose to inject context through concatenation along the frame dimension (shown in Fig. 2), which can flexibly support variable-length context conditions. Specifically, the clean context latents $\mathbf{z}^c$ participate equally with the noisy predicted latents $\mathbf{z}_t$ in the attention computation within DiT Blocks. During output, we only update the noisy latents $\mathbf{z}_t$ using the predicted noise $\boldsymbol{\epsilon}_\phi(\{\mathbf{z}_t, \mathbf{z}^c\}, \mathbf{p}, t)$ while keeping the clean context latents $\mathbf{z}^c$ unchanged.

Another challenge is how to handle positional encoding along the frame dimension in video diffusion models after context frame expansion. Since our method is based on a pre-trained full-sequence text-to-video model, to preserve the original model's generation capability and facilitate easier adaptation to the context-conditioned generation setting, we maintain the same positional encoding for predicted latents $\mathbf{z}_t$ as in the pre-training phase, while assigning new positional encodings to the newly conditioned context latents $\mathbf{z}^c$. Our base model employs RoPE [37], which can conveniently adapt to variable-length position encodings.

3.3. Memory Retrieval

As analyzed in Sec. 1, including all context frames in computation is impractical due to computational overhead and may introduce irrelevant information that causes interference. A reasonable approach is to filter out valuable frames from the context, specifically frames that share overlapping visible regions with the frames to be generated. To this end, we propose **Memory Retrieval** to accomplish this task as shown in Fig. 3 (a). Below, we first introduce several alternative implementation methods, followed by our solution.

Alternative method #1: random selection. A baseline randomly selects frames from context. This works well in early generation when context size is small, as adjacent frames' natural redundancy reduces the risk of missing im-

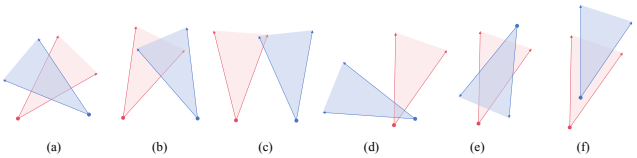


Figure 4. **Examples of FOV Overlap.** We simplify FOV overlap detection to checking intersections between four rays from two camera origins. A practical rule that works for most cases requires: both left and right ray pairs intersect (a, b). However, we must filter out cases where intersection points are either too near (d) or too distant (c) from cameras. While this rule may not cover all scenarios and some corner cases exist (e, f), occasional missed or incorrect candidates don't substantially affect overall performance.

portant information. However, with hundreds of context frames, random selection fails to identify valuable frames.

Alternative method #2: neighbor frames within a window. Another approach selects consecutive recent frames within a window near the current predicted frames. While common in existing methods [5, 34, 52], this has key limitations. First, adjacent frames' redundancy means multiple consecutive frames add little new information beyond the most recent frame. Second, ignoring temporally distant frames prevents awareness of previously seen scenes, leading to continuous generation of new scenes and ultimately breaking scene consistency.

Alternative method #3: hierarchical compression. FramePack [54] proposes a hierarchical compression method for context frames into a minimal set (e.g., 2-3 frames). For two-frame case, it allocates space proportionally: the recent frame gets one full frame, the second recent gets half, the third gets a quarter, and so on, totaling two frames. While achieving high compression, this exponential decay significantly loses historical information. Though the authors suggest manually preserving certain key frames uncompressed, they don't specify the selection criteria.

Our method: camera-trajectory-based search. The fundamental limitation of these methods lies in their inability to identify truly valuable frames from the large number of context frames. They either introduce many redundant frames or lose too much useful information, especially from the old frames that are temporally distant. We leverage the known camera trajectory of the context to search for valuable frames, specifically those that share high-overlap visible regions with predicted frame as shown in Fig. 3 (b).

The first question is how to obtain the camera trajectory of the context video. Since we have introduced camera control into our video generation model in Sec. 3.1, these context frames are generated with user-provided camera poses. These conditioning camera poses can serve as camera annotations for the generated context, eliminating the need for an additional camera pose estimator.

The second question is how to determine co-visibility between frames given their camera poses. We determine this by checking if there is an overlapping region between the fan-shaped areas corresponding to the Fields of View (FOV) of the two cameras. Specifically, since we restrict camera movement to the XY plane, we only need to consider the left and right rays shooting from each camera's origin. By checking the intersection of these four rays from two cameras, we can quickly determine the FOV overlap as shown in Fig. 4. Additionally, we calculate the distance between the predicted frame's camera and the calculated intersection points to eliminate cases where the cameras are too far apart (which typically indicates no actual overlap or very small overlap). This FOV overlap detection is not perfect, as it may fail in cases with occlusions. However, this method effectively reduces the number of candidate context frames.

The final question is: after FOV co-visibility filtering, if the number of filtered frames still exceeds the context condition limit, how should we further filter them? A baseline

Algorithm 1: Training Process of **Context-as-Memory**

Input: Video sequence $\mathcal{X}$ and camera annotations $\mathcal{C}$ in training dataset, context size k

while not converged **do**
- Randomly select predicted video sequence $\mathbf{x}_0$ from $\mathcal{X}$;
- Retrieve k frames as context $\mathbf{x}^c$;
- Obtain camera poses $\{\mathbf{cam}_0, \mathbf{cam}^c\}$ for $\{\mathbf{x}_0, \mathbf{x}^c\}$ from $\mathcal{C}$;
- Obtain latent embeddings $\{\mathbf{z}_0, \mathbf{z}^c\} \leftarrow \text{Encoder}(\{\mathbf{x}_0, \mathbf{x}^c\})$;
- Sample $t \sim U(1, T)$ and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, then corrupt $\mathbf{z}_0$ to $\mathbf{z}_t$;
- Train $p(\mathbf{z}_{t-1} \mid \mathbf{z}_t, \mathbf{z}^c, \mathbf{cam}_0, \mathbf{cam}^c, t)$ using diffusion loss;

Algorithm 2: Inference Process of **Context-as-Memory**

Input: Initial frame set $\mathcal{X} = \{\mathbf{x}_{\text{init}}\}$ and camera poses $\mathcal{C} = \{\mathbf{cam}_{\text{init}}\}$

Output: Generated video sequence $\mathcal{X}$

while generation not finished **do**
- User provides next target camera pose $\mathbf{cam}^t$;
- Retrieve context frames $\mathbf{x}^c \subset \mathcal{X}$ and $\mathbf{cam}^c \subset \mathcal{C}$ by checking FOV overlap with $\mathbf{cam}^t$;
- Compute context latent $\mathbf{z}^c \leftarrow \text{Encoder}(\mathbf{x}^c)$;
- Sample noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and infer latent $\mathbf{z}^t \sim p(\mathbf{z}^t \mid \epsilon, \mathbf{z}^c, \mathbf{cam}^t, \mathbf{cam}^c)$;
- Decode generated frames $\mathbf{x}^t \leftarrow \text{Decoder}(\mathbf{z}^t)$;
- Append $\mathbf{x}^t$ to $\mathcal{X}$ and $\mathbf{cam}^t$ to $\mathcal{C}$;

approach would be random selection, but we also provide some more insightful strategies: (1) Considering the redundancy between adjacent frames, we randomly select only one frame from each group of consecutive frames in the filtered context. This design is highly effective, significantly reducing the number of candidate frames while preserving most of the valuable information. (2) Building upon the first strategy, we can additionally select a few context frames that are furthest apart either spatially or temporally. This helps to supplement potentially missing long-term information (both spatial and temporal). However, in most cases, this additional selection may not be necessary.

Implementation details in training and inference. Assume the maximum number of retrieved context frames is k. During training, we read a long ground truth video (containing thousands of frames) and randomly select a segment as the sequence to be predicted. We then apply our Memory Retrieval method to select $k-1$ context frames from the remaining frames. The overlapping relationships between frames have been pre-computed, eliminating the need for repeated calculations. The first frame of the prediction sequence is also included as an additional context frame to ensure video continuity. Additionally, there is a 10% probability during training that only the recent context frame is used, simulating the beginning of long video generation where no context frames are available. During inference, for each video segment to be predicted, we search $k-1$ context frames from the previously generated frames using FOV-based Memory Retrieval and add the most recently generated frame to the context. The training and inference procedures are outlined in Algorithm 1 and 2, respectively.

3.4. Data Collection

To validate our method, we require long video datasets with camera pose annotations. However, currently available datasets with camera pose information typically consist of short video clips [2, 56]. To obtain long-duration data with precise camera annotations, we utilized a simulation environment, specifically Unreal Engine 5. We generated randomized camera trajectories navigating through different scenes and rendered corresponding long videos. Our dataset comprises 100 videos of 7,601 frames each, featuring 12 distinct scene styles, with captions annotated by a multimodal LLM [48] every 77 frames. To simplify the problem while still effectively validating our method, we constrained the camera trajectory's position changes to a 2D plane and limited rotation to only around the z-axis, which still presents sufficient complexity for camera trajectory control. Additional details about the dataset are provided in the supplementary materials.

4. Experiments

4.1. Experiment Settings

Implementation Details. Our method is implemented on an internal 1B-parameter pre-trained text-to-video Diffusion Transformer, developed for research purposes. The resolution of generated videos is 640×352. The model supports generation of 77-frame videos, with a temporal compression ratio of 4 in the causal 3D VAE, resulting in 20-frame video latents generation. We set the context size to 20, meaning 20 RGB frames are selected as context. Since these frames lack temporal continuity, they are individually compressed using the causal 3D VAE, also resulting in 20 frames of video latents. The model was trained on our collected dataset for over $10,000$ iterations with a batch size of 64 on 8 NVIDIA A100 GPUs. During sampling, we employ Classifier-Free Guidance [15] for text prompts, with 50 sampling steps.

Evaluation Methods. To evaluate our method, we held out 5% of the dataset containing diverse scenes for testing. Our evaluation metrics include: (1) **FID and FVD** for video quality assessment; (2) **PSNR and LPIPS** for quantifying memory capability through pixel-wise differences between frames. Given the lack of memory evaluation methods, we propose two approaches: (1) **Ground truth comparison**: evaluating whether predicted frames match ground truth based on context selected from ground truth frames; (2) **History context comparison**: comparing newly generated frames with previously generated ones in long video sequences. This second approach provides stronger evidence of memory capability as it evaluates consistency in newly generated content. In our implementation, we test on simple trajectories where the camera rotates n degrees and returns, allowing easy identification of corresponding frames for PSNR/LPIPS calculation.

4.2. Comparison Results

In this section, we evaluate memory capabilities across baseline methods, SOTA approaches, and our Context-as-Memory. The compared methods include: (1) Single-frame context using the first frame; (2) Multi-frame context using the first frame plus random historical frames; (3) Diffusion Forcing Transformer (DFoT) [34], using a fixed-size window of most recent frames; (4) FramePack [54], which hierarchically compresses previous context into two frames, with each frame's height or width halved compared to its predecessor. While theoretically supporting all historical frames, compression becomes impractical after several frames as latent size reduces to 1×1. For fair comparison, all methods were implemented on our base model and dataset with identical training configurations and iterations. Results are presented in Tab. 1 and Fig. 5.

Methods	Ground Truth Comparison				History Context Comparison			
	PSNR↑	LPIPS↓	FID↓	FVD↓	PSNR↑	LPIPS↓	FID↓	FVD↓
1st Frame as Context	15.72	0.5282	127.55	937.51	14.53	0.5456	157.44	1029.71
1st Frame + Random Context	17.70	0.4847	115.94	853.13	17.07	0.3985	119.31	882.36
DFoT [34]	17.63	0.4528	112.96	897.87	15.70	0.5102	121.18	919.75
FramePack [54]	17.20	0.4757	121.87	901.58	15.65	0.4947	131.59	974.52
Context-as-Memory (Ours)	**20.22**	**0.3003**	**107.18**	**821.37**	**18.11**	**0.3414**	**113.22**	**859.42**

Table 1. **Quantitative Comparison results.** Due to learning abundant context, Context-as-Memory demonstrates the best memory and highest quality of generated videos. In contrast, DFoT [34] and FramePack [54], which can only utilize the most recent contexts, show relatively inferior performance, even worse than random context selection. This is because although random selection cannot guarantee to select useful information, on average it tends to obtain more information compared to methods that only learn from the most recent context.

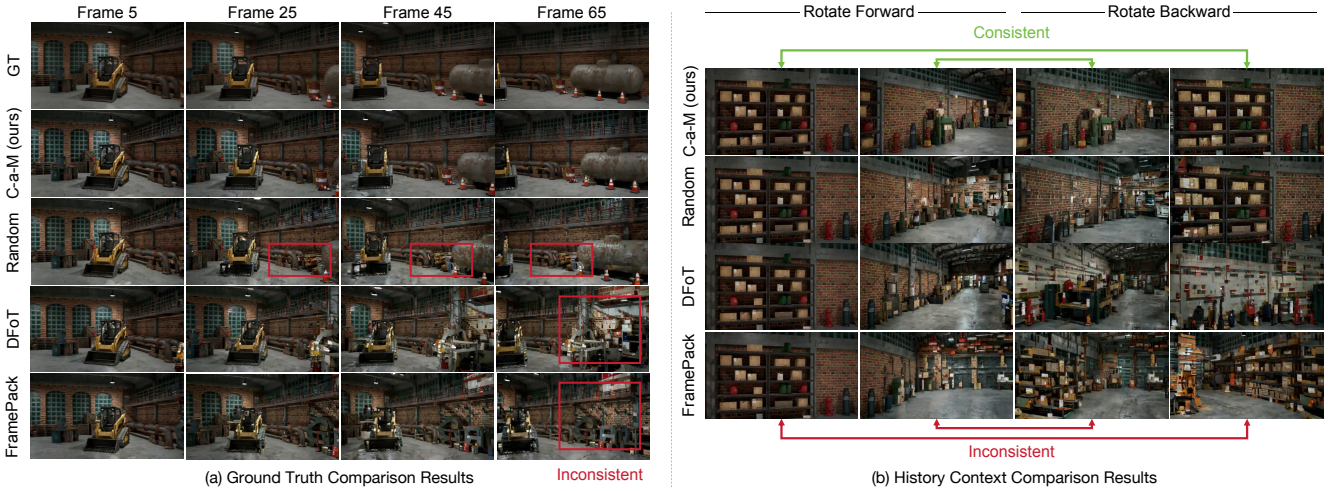


Figure 5. **Qualitative Comparison Results.** Context-as-Memory demonstrates the best memory and the highest visual quality, indicating the effectiveness of sufficient context conditioning. Other methods exhibit scene inconsistency issues due to limited context utilization.

PSNR and LPIPS metrics demonstrate our Memory-as-Context's advantages over other approaches. It effectively retrieves and utilizes useful context information, while other methods have limited context access. Random context selection outperforms DFoT and FramePack, possibly because although it cannot guarantee selecting useful context, it still performs better on average than methods limited to only recent frames. DFoT and FramePack's performance limitations stem from adjacent frame redundancy. Despite accessing dozens of recent frames, the inherent redundancy limits effective information utilization. FramePack's exponential information decay further weakens its memory capabilities compared to DFoT.

Moreover, FID and FVD show our Context-as-Memory achieves the best generation quality among all methods. Sufficient context conditioning not only enhances memory but also improves generation quality by reducing error accumulation in long videos. This improvement stems from two factors: (1) context provides stronger conditional guidance by reducing generation uncertainty, and (2) earlier generated frames used as context contain fewer accumulated errors, helping minimize error propagation in new frames.

Additionally, History Context Comparison proves more challenging than Ground Truth Comparison. Even with simple "rotate forward and rotate backward" trajectories, the performance gaps between methods are significant. DFoT and FramePack can only utilize the most recent context, causing them to continuously generate new content. Only by having access to global context and extracting useful relevant information from it can memory-aware new video generation be achieved.

4.3. Ablation Study

Ablation of Context Size. We studied how context size affects memory capability. Larger contexts theoretically provide more useful information, improving memory performance as shown in Tab. 2. However, this comes with increased computational cost and slower generation speed. When context size reaches 30, there's a notable speed drop compared to size 1. Balancing performance and speed, a

Context Size	GT Comp. PSNR↑	GT Comp. LPIPS↓	HC Comp. PSNR↑	HC Comp. LPIPS↓	FPS↑
1	15.72	0.5282	14.53	0.5456	1.60
5	17.37	0.4825	15.97	0.5063	1.40
10	19.14	0.3554	17.75	0.3985	1.20
20	20.22	0.3003	18.11	0.3414	0.97
30	20.31	0.3137	18.19	0.3319	0.79

Table 2. **Ablation of Context Size.** Larger context contain more useful information and lead to better memory, but also incur higher computational overhead, necessitating an optimal trade-off choice.

Strategy	GT Comp. PSNR↑	GT Comp. LPIPS↓	HC Comp. PSNR↑	HC Comp. LPIPS↓
Random	17.70	0.4847	17.07	0.3985
FOV+Random	19.17	0.3825	17.47	0.3896
FOV+Non-adj	20.11	0.3075	**18.19**	0.3571
FOV+Non-adj+Far-st	**20.22**	**0.3003**	18.11	**0.3414**

Table 3. **Ablation of Memory Retrieval Strategy.** The filtering methods of "FOV" and "Non-adj" (where only one frame from continuous frame sequences is selected as a candidate) effectively filter out useless and redundant information, leading to significant improvements in memory capability.

context size of 20 offers a good trade-off. Future improvements in context compression techniques may help reduce the optimal context size further.

Ablation of Memory Retrieval Strategy. We ablated different memory retrieval strategies to analyze their effects. "Random" refers to randomly selecting context; "FOV+Random" means first filtering using the FOV-based method, then randomly selecting from the remaining candidates; "Non-adj" means only one frame from continuous frame sequences will be selected as a candidate; "Far-st" means frames that are more distant in time or space are more likely to be selected. The results in Tab. 3 demonstrate the effectiveness of "FOV" and "Non-adj" methods in removing useless and redundant information, which significantly increases the probability of selecting useful context and thereby enhances memory capability. The impact of "Far-st" is relatively minor.

4.4. Open-Domain Results

Due to our diverse training dataset and the various visual priors learned by our base model during pre-training, our method has the potential to generalize to open-domain scenarios not present in the training set. We selected images of different styles from the internet and used them as the first frame to generate long videos. We validated using the trajectory of "rotate away and rotate back," which is suitable for verifying memory consistency in generated content. Results in Fig. 6 demonstrate that our method indeed possesses good memory capability in open-domain scenarios.

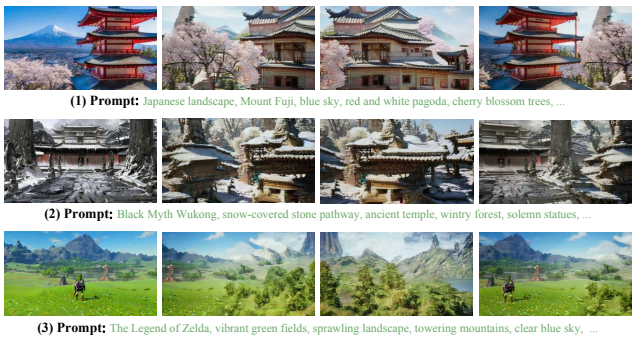


Figure 6. **Open-Domain Results.** We collected open-domain images from the internet and used them as the first frame to generate subsequent long videos. Under the trajectory of "rotate away and rotate back," even when generating new content, it still demonstrates good memory capability.

5. Conclusion

In this work, we propose **Context-as-Memory**, highlighting that using historical generated frames as memory is key to achieving scene-consistent long video generation. Our method design is simple yet effective, directly saving context frames as memory and inputting the context together with the predicted frame as conditions. Furthermore, to avoid high computational overhead caused by lengthy context, we propose **Memory Retrieval** to dynamically select truly valuable context based on the predicted video frames.

Limitations and Future Work. Although our method has made significant progress in achieving memory capability for long video generation, several limitations remain: (1) Our method is limited to static scenes, while memory retrieval for dynamic scenes poses greater challenges; (2) In complex scenarios, particularly those with multiple occlusions (e.g., interconnected indoor rooms), FOV overlap may struggle to effectively identify truly relevant context frames; (3) The inherent error accumulation problem in long video generation persists, which currently can only be addressed through larger datasets, more extensive training, and more powerful base models. In the future, we will continue to develop memory capabilities for open-domain long video generation on larger-scale base models, supporting more complex trajectories, broader scene ranges, and longer generation sequences.

References

[1] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints, 2024. 3

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647, 2025. 3, 4, 6

[3] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. arXiv preprint arXiv:2405.04233, 2024. 3

[4] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. arXiv preprint arXiv:2407.01392, 2024. 2, 3

[5] Etched Decart. Oasis: A universe in a transformer. https://oasis-model.github.io/, 2024. 2, 3, 5

[6] Google DeepMind. Genie 2: A large-scale foundation world model. https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/, 2024. 2, 3

[7] Google DeepMind. Veo 2: Our state-of-the-art video generation model. https://deepmind.google/technologies/veo/veo-2/, 2024. 3

[8] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169, 2024. 2, 3

[9] 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. 3

[10] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. In ICLR, 2025. 3

[11] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. arXiv preprint arXiv:2405.17398, 2024. 2

[12] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325, 2025. 2, 3

[13] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. arXiv preprint arXiv:2503.10589, 2025. 3

[14] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101, 2024. 3

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022. 6

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. Advances in neural information processing systems, 2020. 2

[17] 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. 2

[18] Anssi Kanervisto, Dave Bignell, Linda Yilin Wen, Martin Grayson, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Tabish Rashid, Tim Pearce, Yuhan Cao, et al. World and human action models towards gameplay ideation. Nature, 638(8051):656–663, 2025. 2, 3

[19] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 3

[20] Kling. Kling ai: Next-generation ai creative studio. https://app.klingai.com/, 2024. 3

[21] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José 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. arXiv preprint arXiv:2312.14125, 2023. 2, 3

[22] 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. 2, 3

[23] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. arXiv preprint arXiv:2406.11838, 2024. 3

[24] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022. 2

[25] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022. 2

[26] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. You see it, you got it: Learning 3d creation on pose-free videos at scale. arXiv preprint arXiv:2412.06699, 2024. 3

[27] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In Proceedings of the AAAI conference on artificial intelligence, 2024. 4

[28] OpenAI. Creating video from text. https://openai.com/index/sora/, 2024. 2, 3

[29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023. 3

[30] Yiran Qin, Zhelun Shi, Jiwen Yu, Xijun Wang, Enshen Zhou, Lijun Li, Zhenfei Yin, Xihui Liu, Lu Sheng, Jing Shao, et al. Worldsimbench: Towards video generation models as world simulators. arXiv preprint arXiv:2410.18072, 2024. 2

[31] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. arXiv preprint arXiv:2503.03751, 2025. 3

[32] Runway. Runway : Tools for human imagination. https://runwayml.com/, 2024. 2, 3

[33] Lloyd Russell, Anthony Hu, Lorenzo Bertoni, George Fedoseev, Jamie Shotton, Elahe Arani, and Gianluca Corrado. Gaia-2: A controllable multi-view generative world model for autonomous driving. arXiv preprint arXiv:2503.20523, 2025. 2

[34] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. arXiv preprint arXiv:2502.06764, 2025. 2, 3, 5, 6, 7

[35] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. Advances in neural information processing systems, 2019. 2

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. International Conference on Learning Representations, 2021. 2

[37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. Neurocomputing, 568:127063, 2024. 4

[38] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. arXiv preprint arXiv:2408.14837, 2024. 2, 3

[39] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025. 2, 3

[40] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869, 2024. 2, 3

[41] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In ACM SIGGRAPH 2024 Conference Papers, 2024. 3, 4

[42] Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369, 2025. 3

[43] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors, 2023. 4

[44] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157, 2021. 3

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

[46] Sherry Yang, Jacob C Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In Proceedings of the 41st International Conference on Machine Learning, 2024. 2

[47] 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, 3

[48] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024. 6

[49] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6658–6667, 2024. 3

[50] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video. arXiv preprint arXiv:2504.21853, 2025. 2

[51] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Position: Interactive generative video as next-generation game engine. arXiv preprint arXiv:2503.17359, 2025. 2

[52] Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos, 2025. 2, 3, 5

[53] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048, 2024. 3

[54] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. arXiv preprint arXiv:2504.12626, 2025. 2, 3, 5, 6, 7

[55] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF international conference on computer vision, pages 3836–3847, 2023. 4

[56] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817, 2018. 6

6. Introduction of the Base Text-to-Video Generation Model

We use a transformer-based latent diffusion model as the base T2V generation model, as illustrated in Fig. 7. We employ a 3D-VAE to transform videos from the pixel space to a latent space, upon which we construct a transformer-based video diffusion model. Unlike previous models that rely on UNets or transformers, which typically incorporate an additional 1D temporal attention module for video generation, such spatially-temporally separated designs do not yield optimal results. We replace the 1D temporal attention with 3D self-attention, enabling the model to effectively perceive and process spatiotemporal tokens, thereby achieving a high-quality and coherent video generation model. Specifically, before each attention or feed-forward network (FFN) module, we map the timestep to a scale, thereby applying RMSNorm to the spatiotemporal tokens.

7. Details of Collected Dataset

In this section, we provide a detailed description of the rendered dataset used to train our model.

3D Environments We collect 12 different 3D environments assets from https://www.fab.com/. To minimize the domain gap between rendered data and real-world videos, we primarily select visually realistic 3D scenes, while choosing a few stylized or surreal 3D scenes as a supplement. To ensure data diversity, the selected scenes cover a variety of indoor and outdoor settings, such as city streets, shopping malls, and the countryside.

Camera Trajectories To create data that roam within a scene, we employ smoothed polylines as camera trajectories. Specifically, we begin by randomly sampling coordinate points in the 3D scene to serve as the endpoints of the polyline, and then generate B-spline curves from these points. To ensure smooth camera movement without abrupt speed changes or rotations, we limit the camera's movement distance to the range of [3m, 6m] for each 77-frame video segment and restrict the rotation angle within the xy-plane to less than 60 degrees.

Upon completing the 3D scene collection and trajectory design, we utilized Unreal Engine 5 to batch-render 100 long videos for training. Each video features 7,601 frames (30 fps) of continuous camera movement. Additionally, we record the camera's extrinsic and intrinsic parameters for each frame. The camera is configured with a focal length of 24mm, an aperture of 10, and a field of view (FOV) of 52.67 degrees.

8. Additional Open-Domain Results

In Fig. 8 and Fig. 9, we present additional open-domain results. Using diverse images collected from the internet as initial frames, we demonstrate long video generation with "rotate away and rotate back" trajectories. These source images, representing various styles and scenes, can be found in the provided Data.

Our method achieves generalization capability in open-domain scenarios due to two main factors: (1) Training on diverse scenes enables the model to develop generalizable context utilization skills; (2) The pre-trained base model possesses strong generative priors from exposure to various data types during pre-training.

However, our method still faces significant limitations in open-domain generalization that require future research: (1) The 1B-parameter base model's capabilities are insufficient, only showing good results on simple trajectories. For complex trajectories, the base model struggles to generate high-quality content from the initial frame, leading to unacceptable error accumulation in long video generation. Validating our approach with larger-scale base models remains a future research direction. (2) The method cannot yet support more complex, diverse, and dynamic long-term scene exploration in open-domain settings. Our ideal goal is to enable free, extended navigation from any given image while maintaining memory consistency. This is a challenging objective, though the "context as memory" concept shows promise.

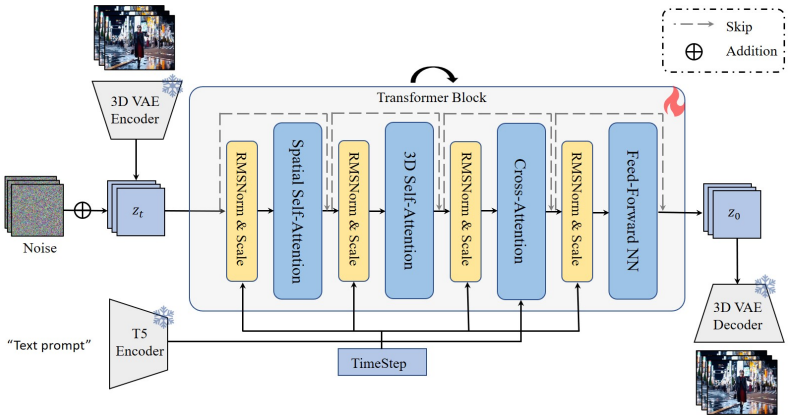


Figure 7. Overview of the base text-to-video generation model.

(1) Prompt: Japanese landscape, Mount Fuji, blue sky, red and white pagoda, cherry blossom trees, cultural harmony, natural beauty.

(2) Prompt: Black Myth Wukong, snow-covered stone pathway, ancient temple, wintry forest, towering trees, solemn statues, sacred ambiance, icicles, ...

(3) Prompt: Black Myth Wukong, mist-laden forest, weathered wooden structure, overgrown vegetation, moss-covered stone steps, ornate tiled roof, ...

(4) Prompt: Black Myth Wukong, a barren landscape of red and grey features a desolate shrine with weathered East Asian architecture, a lone stone statue, ...

(5) Prompt: A sunlit alpine valley shows logs by a rustic cabin, surrounded by meadows, forests, and towering mountains. The breeze gently stirs the grass, ...

(6) Prompt: The Legend of Zelda, vibrant green fields, wildflowers, sprawling landscape, winding rivers, distant ruins, towering mountains, clear blue sky, ...

(7) Prompt: Genshin Impact, a festival pier glows with lanterns, featuring a poised girl in traditional attire amid dragon-shaped lanterns and swaying pavilions, ...

(8) Prompt: Genshin Impact, a gentle pan reveals a vibrant countryside village in golden morning light, with a windmill turning above timber-framed houses and ...

Figure 8. **Open-Domain Results.**

(1) Prompt: A smooth tracking shot sweeps through a sunlit high-rise office, showcasing ergonomic workstations and a stunning cityscape view, ...

(2) Prompt: A bright, open-plan office features minimalist desks, ergonomic chairs, and calming greenery, all under natural and LED lighting, ...

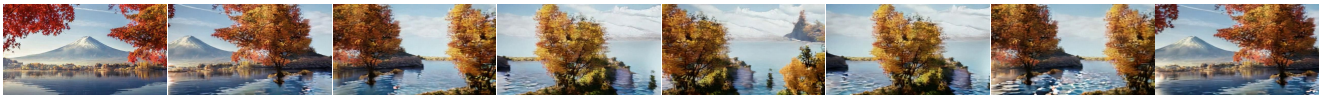
(3) Prompt: A serene bedroom features a neatly made bed with cream bedding, wooden nightstands, and minimalist decor. Soft natural light and framed artwork ...

(4) Prompt: A slow glide through a formal garden reveals a circular stone pedestal, symmetrical hedges, and blooming rose bushes, ...

(5) Prompt: An aerial view glides over a winding path in an alpine meadow, with golden sunlight and mist creating a serene morning scene, ...

(6) Prompt: A slow zoom captures a tranquil alpine forest and a mountain range glowing in the sunset. A still body of water reflects snow-capped peaks ...

(7) Prompt: A sunlit view of a frozen alpine basin surrounded by snow-covered peaks. The camera pans, highlighting the contrast between sunlit ridges ...



(8) Prompt: A tranquil lakeside scene reflects a snow-capped mountain, with golden light revealing shoreline buildings amidst vibrant autumn foliage ...

(9) Prompt: A sunlit mountain path winds through a lush valley with wildflowers and a rustic cabin. Snow-capped peaks rise above, set against a crisp blue sky, ...

(10) Prompt: A sunny alpine path curves along a hillside, bordered by evergreens and a stone wall. A fence overlooks a valley with rooftops, ...

(11) Prompt: In an arid desert under a burnt-orange sky, a dusty six-wheeled rover stands amid scattered components, evoking a Mars expedition base.

Figure 9. **Open-Domain Results.**