VOXELMEM: VIDEO GENERATION WITH 4D VOXEL CONTROL AND MEMORY

Anonymous authors
Paper under double-blind review

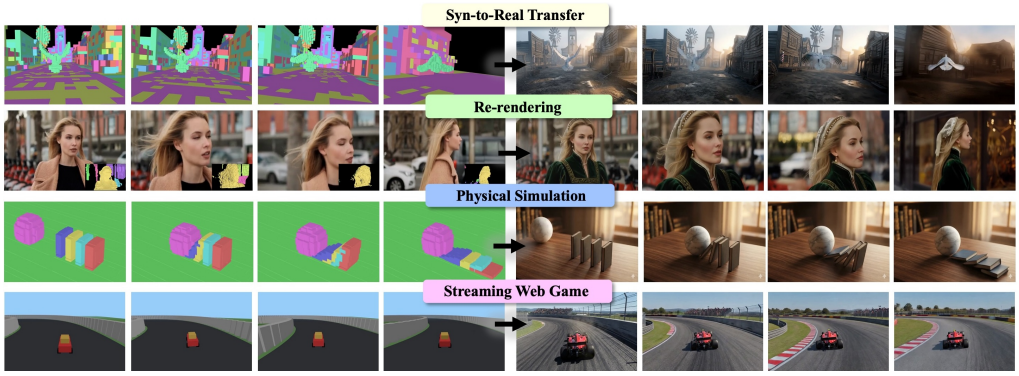


Figure 1: Given 4D voxel controls, VoxelMem generates videos across diverse applications, including filmmaking (e.g., synthesis-to-real transfer and re-rendering), realistic physical simulation, and streaming web-game generation, while persistent voxel memory maintains long-term consistency.

ABSTRACT

Long video generation remains limited by weak controllability, spatial drift, and memory loss when objects leave and re-enter the view. We propose *VoxelMem*, a framework that uses 4D voxels as both controllable primitives and memory carriers. Each voxel stores time-dependent geometry, visibility, segmentation identity, and local appearance, enabling multi-granularity control from object-level layout to part-level motion. To maintain long-term consistency, VoxelMem combines local appearance propagated with voxel motion and global history frames retrieved by future voxel coverage, providing both fine-grained spatial correspondence and holistic scene references. We further extend VoxelMem to streaming generation by distilling the bidirectional model into a causal generator. Finally, we construct 4D video–voxel training data from synthetic rendered scenes and real-world videos to support multi-granularity control and memory learning.

1 INTRODUCTION

Recent advances in video generation models (Yang et al., 2024b; Wan et al., 2025; Xing et al., 2025b) have produced impressive short clips, but long, controllable video generation remains an open problem. Existing methods face two key limitations. First, they rely mainly on global prompts or initial images, limiting fine-grained motion control (Yang et al., 2025; Xiao et al., 2025; Meng et al., 2025). Second, they lack persistent 4D scene memory, causing spatial drift, identity loss, and inconsistent recovery of revisited regions (Yu et al., 2025; Zhu et al., 2025; Sun et al., 2025; Xiao et al., 2026). These limitations call for an explicit spatiotemporal representation that can support both controllable guidance and long-term scene memory. In this work, we introduce 4D voxels as such a representation. Voxels provide a simple and stable interface that is easy to craft, manipulate, and couple with physical simulation, while offering spatially grounded and memory-efficient slots for storing visual information. As illustrated in Fig. 1, this voxel-based representation enables voxel-controlled filming (rows 1–2), simulation-guided generation with physically plausible dynamics (row 3), and streaming web games (row 4).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Early controllable video generators operate in image space, using depth (Lin et al., 2024; Jiang et al., 2025), sketches (Xing et al., 2024; Guo et al., 2024), bounding boxes (Wang et al., 2024; Ma et al., 2024), or segmentation maps (Li et al., 2025a; Xing et al., 2025a) as frame-level guidance, but such image-space controls are limited in specifying 3D motion. Recent methods therefore introduce geometry-aware controls, such as dense point clouds (Yu et al., 2024; Huang et al., 2025b; Ren et al., 2025), sparse trajectories (Gu et al., 2025; Zhang et al., 2025d), 3D bounding boxes (Wang et al., 2025b), or hybrid foreground-background primitives (Zheng et al., 2026). However, each representation has limitations. Dense point clouds provide geometric cues but are noisy, redundant, and difficult to author for dynamic objects. Sparse trajectories offer motion anchors but preserve little appearance or scene information. 3D bounding boxes encode object-level volume but are too coarse for fine-grained motion or deformation. Hybrid primitives (Zheng et al., 2026) partially bridge this gap by combining dense background points with foreground 3D Gaussian spheres, but still rely on predefined primitive shapes that are inflexible across object categories or granularities. These limitations motivate a volumetric, multi-granular control primitive, which our 4D voxels provide.

Beyond control, long video generation requires long-term scene memory, especially when dynamic objects leave and re-enter the field of view or when the camera revisits previously observed regions. Existing memory methods can be broadly grouped into non-geometric and geometry-based approaches. Non-geometric methods compress past context into learned tokens or implicit memory states (Dalal et al., 2025; Zhang et al., 2025a; Cai et al., 2025), but such memories are not spatially grounded and cannot guarantee 3D consistency. Geometry-based methods take either a global or local form: global memory provides scene-level references to recover appearance and correct artifacts, while local memory establishes fine-grained spatial correspondences. Global memory methods select reference frames according to camera poses, 4D occupancy, or related geometric cues (Yu et al., 2025; Huang et al., 2025a; Li et al., 2025b), but keep appearance at the frame level and cannot assign visual information to precise spatial locations. Local memory methods usually update RGB point clouds to guide generation (Huang et al., 2025b; Wu et al., 2026), but attaching features to isolated points often produces fragmented references and loses surrounding image context. Recent methods further explore carrying latent-level features (Yu et al., 2026b; Wang et al., 2026a), yet still suffer from two limitations: the carried patches are tied to the latent grid resolution and are therefore inflexible, while latent patches may contain entangled high-level semantics rather than clean local appearance. In contrast, 4D voxels provide stable spatiotemporal slots with explicit extent: each voxel can store features from a local image patch whose size adapts to the voxel scale, while the voxel field can index history frames by geometric coverage. This allows voxel memory to combine local correspondence with global reference retrieval, reducing fragmented point-based guidance while preserving coherent visual evidence for revisited regions and moving objects.

We realize these ideas in VoxelMem, a long video generation framework that uses 4D voxels as a unified representation for control and memory. It maintains a dynamically updated voxel field, where each voxel records its time-dependent position, segmentation identity, local image patch feature, and visibility. For future generation, the propagated voxel field provides two types of conditioning. The control path projects voxels into a color-coded canvas based on segmentation identity and injects its encoded latent through lightweight channel concatenation to guide layout and motion. The memory path combines local patch features splatted by propagated voxel geometry with globally retrieved history frames and injects them through sequence-level token concatenation. This design enables the video generator to use explicit voxel control, geometrically aligned local appearance, and holistic frame-level references within a single 4D representation.

We further extend VoxelMem to streaming generation for interactive applications. While bidirectional video diffusion models provide strong generation quality, their full-sequence denoising process is slow, making them unsuitable for real-time interactive applications such as web games. To address this, we distill the bidirectional model into a causal generator following the self-forcing training strategy of LongLive (Yang et al., 2025), enabling autoregressive blockwise generation with cached historical states. We further adopt relative RoPE (Yi et al., 2025) to improve temporal generalization over long videos. To adapt sequence-concatenated local voxel memory to causal inference, we propose an interleave-forcing scheme, where denoising tokens in each block attend only to the corresponding local memory tokens and past clean video tokens.

Training such a model requires paired video–voxel data at multiple granularities. We construct training pairs from both synthetic rendered scenes and real-world videos: synthetic data provides accurate 4D voxel fields but lacks real-world appearance diversity, while real-world data provides

realistic appearance but requires estimating the underlying 4D structure. For synthetic data, we collect static and dynamic scenes and use Blender as the renderer: static meshes are voxelized using an AABB-based grid, while dynamic objects are converted into temporally consistent voxel trajectories through spatial hashing. For real videos, we first estimate a 4D scene field with long-range tracking (Xiao et al., 2024) and obtain segmentation maps (Ravi et al., 2024) to separate different objects and regions. We then cluster tracked points with similar motion patterns within each segment, producing voxel-like primitives at different granularities. Both sources are normalized into the same 4D voxel format, allowing the model to learn realistic appearance and motion from real footage and accurate controllable geometry from synthetic data.

Our main contributions are:

- We introduce *VoxelMem*, a video generation framework that uses multi-scale 4D voxels for both explicit control and long-term scene memory.
- We propose a hybrid voxel memory that combines propagated local appearance features with globally retrieved history frames for consistent long video generation.
- We extend voxel memory to streaming generation and propose interleave-forcing to make local memory injection compatible with causal inference.
- We develop a 4D video–voxel data construction method using synthetic rendered scenes and real-world videos.

2 Related Work

2.1 Controllable Video Generation

Controllable video generation uses diverse conditions. Text, image, and camera guidance provide semantic, appearance, and viewpoint control (Chen et al., 2024; Guo et al., 2023; Blattmann et al., 2023; Yang et al., 2024b; Wan et al., 2025; Yang et al., 2024a; Yu et al., 2024), but offer limited control over detailed object motion. Image-space controls such as boxes, masks, and depth improve spatial controllability (Li et al., 2025a; Xing et al., 2025a; Bhat et al., 2024; Liu et al., 2026; Lin et al., 2024; Jiang et al., 2025), yet remain frame-wise and provide limited long-range correspondence.

Recent works therefore explore geometry-aware controls. Dense point clouds provide explicit correspondences but are noisy and redundant (Bian et al., 2025a; Huang et al., 2025b; Ren et al., 2025); sparse trajectories offer intuitive motion control but carry little appearance information (Wu et al., 2024; Zhang et al., 2025b; Geng et al., 2025; Burgert et al., 2025; Gu et al., 2025; Zhang et al., 2025d); meshes and 3D boxes provide explicit geometry but are either costly to obtain or too coarse for fine-grained deformation (Liu et al., 2025; Zhang et al., 2025c; Wang et al., 2025b). Hybrid representations combine background points with foreground Gaussian trajectories (Zheng et al., 2026), but rely on separate primitives for different scene components. In contrast, our 4D voxels provide a unified volumetric primitive for both control and memory.

2.2 Memory for Long Video Generation

Long video generation relies on memory to preserve identity, layout, and visual details over extended horizons. Existing methods explore learnable or compressed temporal memory (Dalal et al., 2025; Yu et al., 2026a), through mechanisms such as sparse attention routing (Cai et al., 2025), frame packing (Zhang & Agrawala, 2025), learned history embeddings (Zhang et al., 2025a), and KV memory management (Yi et al., 2026). While these approaches improve long-context efficiency and mitigate temporal drift, their memory is primarily organized in token or temporal space, without explicit correspondence to persistent 3D locations. This makes the stored information less spatially interpretable and provides limited guarantees of 3D consistency.

Another line of work introduces geometry-aware memory for long-range consistency. Global retrieval methods select history frames using camera pose, depth, or geometric overlap (Yu et al., 2025; Huang et al., 2025a; Li et al., 2025b), but retain them as global references without location-specific appearance hints. Local methods establish 3D correspondences through point-cloud memory (Huang et al., 2025b; Wu et al., 2026), yet mainly target static scene regions. Recent latent-memory methods store feature-level evidence (Yu et al., 2026b; Wang et al., 2026a), but their mem-

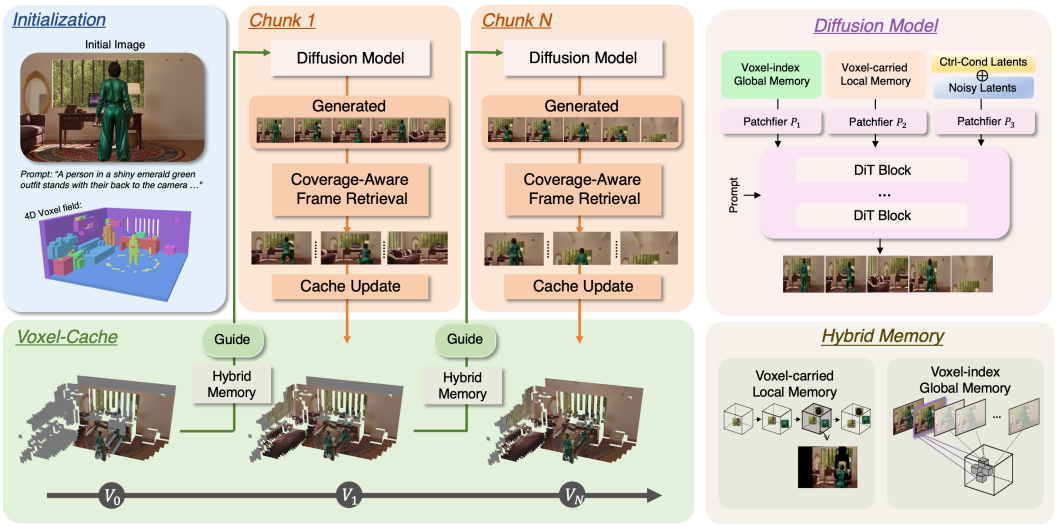


Figure 2: Overview of the VoxelMem framework. Given an initial image and 4D voxel controls, VoxelMem generates videos chunk by chunk while updating a voxel cache. For each target chunk, the voxel field provides voxel-indexed global memory for retrieving relevant frames and voxel-carried local memory for propagating patch-level appearance. Both memory sources are injected into video model to improve long-range consistency under camera revisits and dynamic motion.

ory is tied to latent features rather than explicit spatial units. In contrast, VoxelMem uses persistent 4D voxels to carry local appearance while also indexing globally relevant history frames, supporting both revisited backgrounds and moving objects.

3 Method

3.1 Overview

Given an initial image and a 4D voxel field, VoxelMem generates videos chunk by chunk while maintaining an updated voxel cache. The voxel field provides both control and memory: it guides generation through voxel control, retrieves history frames as voxel-indexed global memory, and propagates local appearance as voxel-carried local memory. Sec. 3.2 defines the voxel primitives, Sec. 3.3 introduces voxel control, Sec. 3.4 presents the hybrid memory scheme, and Sec. 3.5 extends it to causal streaming with interleave-forcing.

3.2 Voxel Primitives

We represent a video with a set of voxel primitives that persist and evolve over time. Formally,

$$\mathcal{V} = \{(\Theta_{t,j}, s_j, \mathcal{M}_j) \mid t = 1, \ldots, T,\ j = 1, \ldots, K\}, \tag{1}$$

where j indexes a persistent voxel, t denotes time, s_j is its segmentation identity, and $\mathcal{M}_j$ is its local memory cache. Each voxel is represented at time t by

$$\Theta_{t,j} = (\mu_{t,j}, d_{t,j}, v_{t,j}, h_{t,j}), \tag{2}$$

where $\mu_{t,j}$ denotes the projected 2D center, $d_{t,j}$ specifies its depth, $v_{t,j}$ indicates visibility, and $h_{t,j}$ denotes the screen-space extent. Together, these time-varying states trace the voxel over time.

3.3 Voxel as Control

We first render the target voxel states into a control condition. For each target frame t, visible voxels are projected onto the image plane to form a color-coded canvas C_t, where each segmentation identity is assigned a unique fixed color and occlusions are resolved by depth order. Then we encode each canvas with the VAE and concatenate it with the noisy latents:

$$\tilde{z}_t = z_t \oplus \mathrm{VAE}(C_t). \tag{3}$$

where $\oplus$ denotes channel-wise concatenation. The concatenated latent $\tilde{z}_t$ is then patchified into denoising tokens X_z, which are later concatenated with memory tokens in Sec. 3.4.

3.4 VOXEL AS HYBRID MEMORY

We build a hybrid memory from the voxel field: global memory M_h retrieves relevant history frames, while local memory M_v propagates voxel-carried appearance to target frames.

3.4.1 VOXEL-INDEXED GLOBAL MEMORY

We retrieve global memory frames by using the voxel field as an index. For a target window $\mathcal{T}$ and history frames $\mathcal{H}$, each candidate frame $\tau \in \mathcal{H}$ covers the future-relevant voxels

$$\mathcal{C}(\tau) = \{j \mid v_{\tau,j} = 1,\ \exists t \in \mathcal{T}, v_{t,j} = 1\}. \tag{4}$$

We select up to K history frames that maximize voxel coverage:

$$\mathcal{H}^* = \arg \max_{\mathcal{S} \subset \mathcal{H}, |\mathcal{S}| \le K} \left| \bigcup_{\tau \in \mathcal{S}} \mathcal{C}(\tau) \right|. \tag{5}$$

The selected frames form the global memory $M_h = \{I_\tau \mid \tau \in \mathcal{H}^*\}$, favoring frames that cover reappearing objects and revisited scene regions. In practice, we additionally apply temporal weighting to favor recent frames while keeping voxel visibility as the primary retrieval criterion.

3.4.2 VOXEL-CARRIED LOCAL MEMORY

In addition to frame-level memory, each voxel maintains a local memory cache

$$\mathcal{M}_j = (p_j, \tau_j), \tag{6}$$

where p_j is the RGBA patch stored by voxel j and τ_j is its source-frame timestamp.

When extracting patches from a history frame, visible voxels are processed from near to far to prevent farther voxels from capturing foreground appearance. Accordingly, $\mathrm{Crop}(I_\tau, \Theta_{\tau,j})$ extracts the projected support of voxel j while excluding pixels already assigned to closer voxels, with the remaining visibility recorded in the alpha channel. The extracted patch is used to update the voxel cache whenever the cache is empty or $|\tau - \tau_j| > \Delta_{\mathrm{mem}}$:

$$\mathcal{M}_j \leftarrow (\mathrm{Crop}(I_\tau, \Theta_{\tau,j}), \tau). \tag{7}$$

The cached patch is then propagated along the trajectory of the same voxel. For each target frame, the warped patches are composited from far to near using depth and alpha, producing the voxel-carried local memory M_v as geometry-aligned appearance guidance.

3.4.3 TOKEN-CONCATENATION INTEGRATION

We inject the history-frame memory M_h and voxel-carried local memory M_v into the video transformer through token concatenation. Let X_z, X_h, and X_v denote the patchified tokens of the noisy latent, history frames, and voxel-carried local memory, respectively. The transformer input is

$$X = [\bar{X}_h; X_z; \bar{X}_v], \tag{8}$$

where $\bar{X}_h$ uses the dedicated temporal RoPE positions, while $\bar{X}_v$ shares the target temporal positions with X_z and applies a spatial offset to distinguish local memory tokens.

3.5 EXTENSION TO STREAMING LONG VIDEO

To support low-latency streaming, we extend VoxelMem to a causal autoregressive model following the self-forcing strategy of LongLive (Yang et al., 2025). The bidirectional generator is distilled with a DMD-based objective, where each block is generated from previously generated video states and supervision is applied only to the current block.

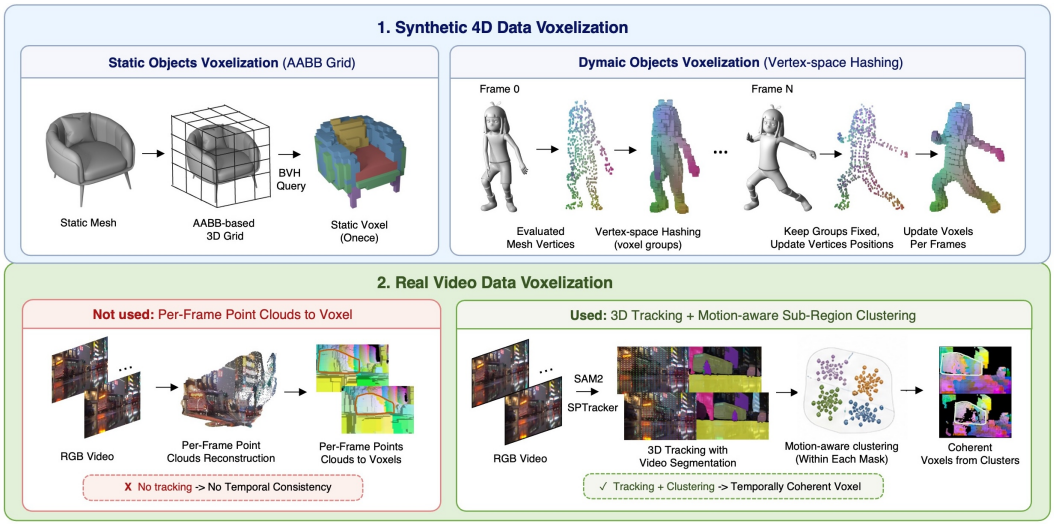


Figure 3: Overview of the video-voxel data construction pipeline. Synthetic scenes provide accurate 4D voxels through static mesh voxelization and dynamic vertex-space hashing. Real videos are converted into voxel-like primitives by combining 3D tracking, segmentation, and motion-aware clustering. Both sources are unified as paired RGB videos and 4D voxels for training.

For voxel-carried local memory, we introduce interleave-forcing. Given a streaming block b with video tokens $\mathbf{x}_b = \{x_{b,i}\}_{i=1}^{F}$ and local-memory tokens $\mathbf{v}_b = \{v_{b,i}\}_{i=1}^{F}$, we interleave them by:

$$\mathbf{s}_b = [x_{b,1}, v_{b,1}, \ldots, x_{b,F}, v_{b,F}]. \tag{9}$$

Since local-memory tokens are frame-specific, they are used only for the current block and are not retained in the streaming cache; instead, we keep a video-only cache, $\mathrm{Cache}_b = \mathrm{KV}(\mathbf{x}_b)$, which avoids unnecessary KV growth while keeping local memory aligned with the current voxel state.

4 Dataset Construction

We construct paired video–voxel data from real-world videos and synthetic rendered scenes. Real videos are converted into voxel-like primitives using segmentation and 3D tracking, while synthetic scenes are constructed in Blender; details are provided in the supplementary material.

5 Experiments

5.1 Experiment Settings

Dataset and Model. For training, we use around 40K real-world videos from VideoPainter (Bian et al., 2025b) and Sekai (Li et al., 2025c), together with 40K CG-synthetic video–voxel pairs. For evaluation, we construct VM-4D, a revisit benchmark covering both same-pose and different-pose revisits. We also evaluate general image-to-video quality on SpatialViD datasets (Wang et al., 2025a) filtered following Versecrafter (Zheng et al., 2026). We train bidirectional Wan2.2 (Wan et al., 2025) models at 5B and 14B scales and further distill the 5B model for streaming generation.

Baselines. We compare with geometry-aware control methods, including FlexTraj (Zhang et al., 2025d) and VerseCrafter (Zheng et al., 2026), and long-video generation methods, including Matrix-Game3 (Wang et al., 2026b), Spmem (Wu et al., 2026), and LiveWorld (Duan et al., 2026).

Metrics. We evaluate revisit consistency using PSNR, SSIM, and LPIPS (Zhang et al., 2018), geometric control using depth RMSE (D-RMSE) (Eigen et al., 2014), and video consistency using foreground/background CLIP-I (Radford et al., 2021). We additionally report VBench-I2V results on eight quality and consistency dimensions, with the overall score computed as their average.

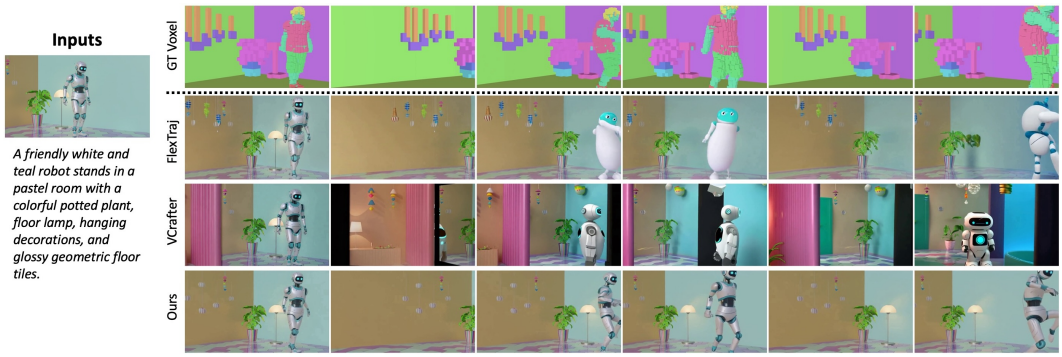


Figure 4: Qualitative comparison with 4D control baselines. Compared with FlexTraj (Zhang et al., 2025d) and VerseCrafter (Zheng et al., 2026), VoxelMem better preserves foreground identity and background when the camera moves away/returns, benefiting from voxel-based hybrid memory.

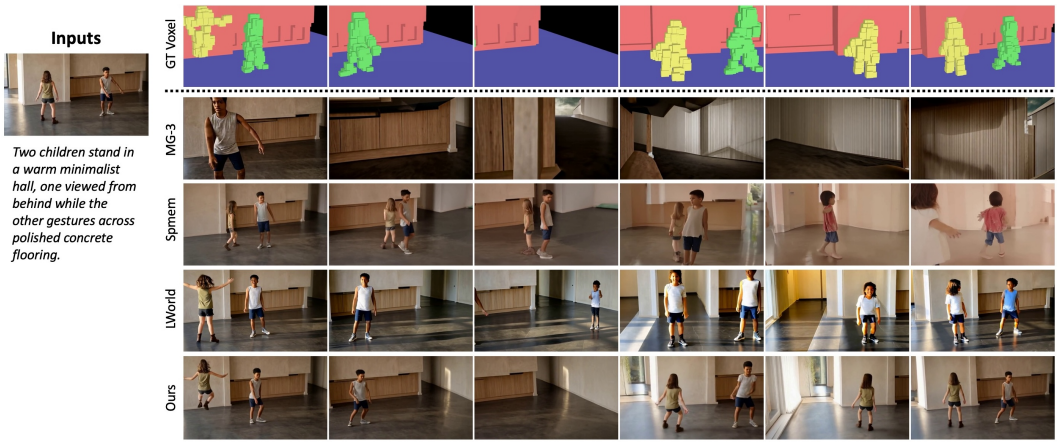


Figure 5: Qualitative comparison with long video generation baselines. MG-3 (Wang et al., 2026b) lacks precise camera and object control, while Spmem (Wu et al., 2026) and LWorld (Duan et al., 2026) struggle with dynamic foregrounds. VoxelMem better preserves scene structure and foreground identity through unified 4D voxel control and memory.

5.2 Qualitative Comparison

Fig. 4 and Fig. 5 compare our method with existing 4D control and long video generation baselines. In Fig. 4, the camera moves out of the scene and later moves back to previously observed regions. FlexTraj (Zhang et al., 2025d) can follow sparse motion cues, but it does not maintain explicit long-term memory. As a result, the robot appearance changes significantly after each revisit. VerseCrafter (Zheng et al., 2026) also suffers from identity drift, and its foreground control is unstable because it represents the foreground object with coarse Gaussian primitives. In contrast, our method stores local appearance in voxel memory and retrieves history frames by future voxel coverage, allowing the robot and the background layout to remain consistent.

Fig. 5 compares long video generation baselines. MG-3 (Wang et al., 2026b) provides interactive control through keyboard actions, but such control does not precisely determine camera speed, target viewpoint, or character pose. It therefore fails to return to the same spatial location. Spmem (Wu et al., 2026) improves long-range generation with memory, but it cannot control foreground objects reliably and its visual quality degrades as the video becomes longer. LWorld (Duan et al., 2026) uses point-cloud memory, but its performance depends on dense and well-aligned point clouds. Even when we lift our voxels into dense point clouds, point-based memory still struggles with unaligned foreground motion and dynamic objects. Our voxel representation provides a more stable control and memory interface, preserving both scene layout and foreground identity during revisit.

Table 1: Quantitative comparison on the VM-4D revisit benchmark. We report bidirectional 5B/14B and distilled 5B streaming models. VoxelMem performs best in both revisit settings, with stronger background recovery, appearance consistency, and control. Second best in orange.

	Same Pose Revisit ($A \to B \to A \to B \to A$)										Different Pose Revisit			
	PSNR$_{bg}$ ↑		SSIM$_{bg}$ ↑		LPIPS$_{bg}$ ↓		D-RMSE↓		CLIP-I↑		D-RMSE↓		CLIP-I↑	
	RetA	RetB	RetA	RetB	RetA	RetB	fg	bg	fg	bg	fg	bg	fg	bg
FlexTraj (5B)	15.90	17.90	0.585	0.575	0.371	0.314	0.228	0.162	0.908	0.908	0.265	0.172	0.849	0.857
VCrafter (14B)	11.94	13.50	0.459	0.416	0.554	0.550	0.360	0.255	0.879	0.817	0.364	0.230	0.823	0.802
Spmem (5B)	13.95	14.73	0.604	0.664	0.544	0.538	0.338	0.211	0.886	0.816	0.371	0.209	0.819	0.807
MG-3 (5B)	13.48	16.26	0.487	0.475	0.484	0.403	0.422	0.246	0.873	0.891	0.479	0.273	0.817	0.830
LWorld (14B)	17.27	14.99	0.566	0.485	0.300	0.374	0.234	0.196	0.917	0.881	0.257	0.191	0.844	0.834
Ours-5B (Streaming)	21.25	21.35	0.654	0.667	0.217	0.189	0.232	0.165	0.934	0.925	0.262	0.169	0.867	0.857
Ours-5B	21.87	24.44	0.697	0.714	0.185	0.140	0.226	**0.161**	0.937	0.942	0.256	**0.166**	**0.869**	0.877
Ours-14B	**24.12**	**26.29**	**0.766**	**0.790**	**0.147**	**0.116**	**0.223**	0.165	**0.939**	**0.944**	**0.253**	**0.166**	**0.869**	**0.881**

Table 2: VBench-I2V evaluation on SpatialViD. Overall is the average of the eight dimensions.

Method	Imaging	Aesthetic	Dynamic	Motion	BG Cons.	Subj. Cons.	I2V BG	I2V Subj.	Overall
FlexTraj (5B)	0.6745	0.5025	**0.6844**	**0.9919**	0.9435	0.9520	0.9865	0.9873	0.8403
VCrafter (14B)	0.7098	0.4958	0.5400	0.9907	**0.9466**	0.9447	0.9864	0.9867	0.8251
LWorld (14B)	**0.7377**	**0.5338**	0.6711	0.9886	0.9233	0.9356	0.9650	0.9695	0.8406
Ours-14B	0.7137	0.4990	**0.6844**	0.9915	0.9429	**0.9532**	**0.9884**	**0.9898**	**0.8454**

5.3 Quantitative Comparison

Table 1 shows that our 14B model achieves the best overall performance across both revisit settings. In same-pose revisits, it substantially improves PSNR to 24.12/26.29 and SSIM to 0.766/0.790 at RetA/RetB, while reducing LPIPS to 0.147/0.116. It also achieves the highest CLIP-I of 0.939/0.944 on foreground/background regions. The 5B bidirectional and streaming variants remain competitive with larger baselines, with the bidirectional 5B model reaching 21.87/24.44 PSNR and 0.937/0.942 CLIP-I. Under different-pose revisits, our method again achieves the best CLIP-I, reaching 0.869/0.881, together with the lowest D-RMSE of 0.253/0.166, showing robust voxel memory.

Table 2 further reports VBench-I2V results on 450 SpatialViD videos. Our method achieves the best overall score (0.8454), outperforming all baselines. It achieves the strongest subject consistency (0.9532) and I2V consistency (0.9884/0.9898 for background/subject), indicating that the proposed voxel-based control and memory improve foreground identity and consistency. The slight gaps in imaging and aesthetic quality may partly stem from the larger proportion of synthetic training data.

5.4 Ablation Study

Memory design. Removing either global or local memory consistently degrades the quantitative results in Table 3 and the revisit quality in Fig. 6. Without global memory, the model lacks holistic historical references, leading to a visible reconstruction artifact, e.g. the region above the person's head. Without local memory, appearance is no longer explicitly anchored to the corresponding voxels, causing the background to change noticeably after revisiting the same region. These results show that global and local memory provide scene-level and voxel-aligned appearance cues.

Condition injection. Replacing sequence concatenation with channel concatenation leads to lower scores in Table 3 and more visible appearance errors in Fig. 6. Channel concatenation mixes memory features with the spatial input early in the network, which can weaken their role as explicit historical references. In contrast, sequence-concatenated memory tokens remain directly accessible through attention, making long-range appearance retrieval more effective.

Memory Encoding. Using latent patches instead of image patches degrades performance in Table 3 and introduces more artifacts in Fig. 6. Latent patches compress richer semantic and contextual information into a lower-resolution representation, making their features less locally aligned with individual image regions. Directly warping such compressed features can therefore propagate mis-

Table 3: Quantitative Ablation Study. All variants use the bidirectional 5B model. Removing either global or local memory degrades performance, while channel-concatenated injection and latent-patch memory are less effective than our default design. The full model performs the best.

	Same Pose Revisit ($A \to B \to A \to B \to A$)										Different Pose Revisit			
	PSNR$_{bg}$ ↑		SSIM$_{bg}$ ↑		LPIPS$_{bg}$ ↓		D-RMSE↓		CLIP-I↑		D-RMSE↓		CLIP-I↑	
	RetA	RetB	RetA	RetB	RetA	RetB	fg	bg	fg	bg	fg	bg	fg	bg
CC-Only Inject	18.92	20.45	0.656	0.629	0.276	0.245	0.285	0.191	0.917	0.915	0.294	0.187	0.856	0.859
w/o GlobalMem	21.08	23.25	0.686	0.701	0.218	0.179	0.228	0.162	0.934	0.933	0.261	0.169	0.864	0.872
w/o LocalMem	17.59	20.61	0.620	0.642	0.308	0.241	0.232	0.161	0.929	0.929	0.271	0.174	0.859	0.861
w/ LatentPatch	18.34	19.83	0.621	0.601	0.290	0.255	0.266	0.179	0.922	0.914	0.291	0.183	0.857	0.857
Ours-5B	**21.87**	**24.44**	**0.697**	**0.714**	**0.185**	**0.140**	**0.226**	**0.161**	**0.937**	**0.942**	**0.256**	**0.166**	**0.869**	**0.877**

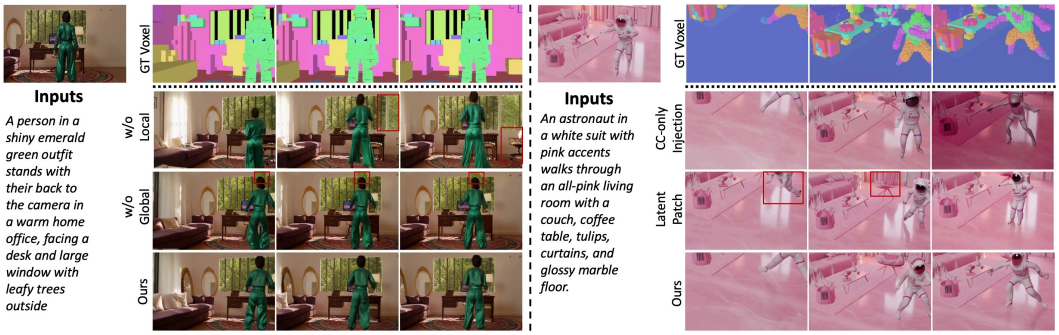


Figure 6: Qualitative Ablation study. Removing either global or local memory weakens appearance preservation and revisit consistency, while channel-concatenated injection and latent-patch memory introduce more visible artifacts. The full design achieves the most stable long-video generation.

matched content and distort local appearance. In contrast, image-space patches retain more explicit pixel-level details and provide a more stable representation for voxel-carried memory.

6 Conclusion and Limitations

We introduced *VoxelMem*, a framework for controllable video generation with 4D voxel control and memory. Voxel primitives encode geometry, visibility, identity, and local appearance, serving as both control signals and persistent memory. Our hybrid memory propagates local appearance while retrieving global history frames, improving consistency under object and scene revisits. We further extend VoxelMem to streaming generation and build a unified real-and-synthetic video-voxel data pipeline. Experiments demonstrate improved long-range consistency and controllability.

VoxelMem still has limitations in handling transparency and voxel merging. Transparent surfaces may entangle voxel memory with background content seen through the surface, while voxel merging does not explicitly propagate appearance changes, e.g., color transfer, across interacting regions. Future work could better model these effects for more robust dynamic scene generation.

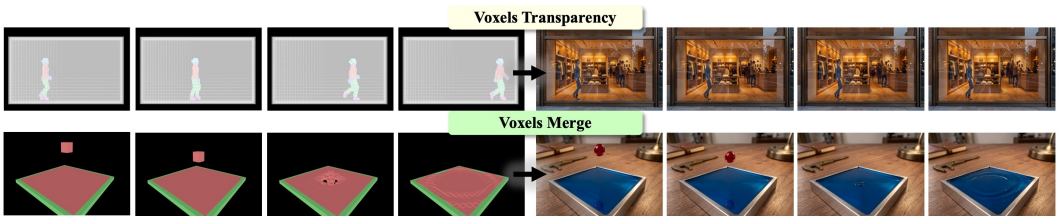


Figure 7: Limitations. Transparent surfaces may entangle memory with background content, suppressing motion behind them. Voxel merging may fail to propagate appearance changes.

AI USE STATEMENT

Large Language Models (LLMs) were used to assist with manuscript writing and the construction of the streaming web-game interface. For writing, we used LLMs to refine language, improve readability, correct grammar, and rephrase sentences throughout the manuscript. These tools were used only for linguistic editing and presentation, and were not involved in formulating the research ideas, methodology, experimental design, or scientific conclusions.

For the streaming web-game application, we additionally used an LLM to generate and refine Three.js code for constructing the interactive voxel-style game interface, including scene layout, game logic, and user controls. The generated code and resulting interactive behaviors were manually inspected and verified by the authors before being used in our demonstrations.

All research concepts, technical designs, experiments, analyses, and conclusions were developed and validated by the authors. The authors take full responsibility for both the manuscript content and any LLM-assisted code used in the project.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. In this study, no human subjects or animal experimentation was involved. All datasets used, including VM-4D, were sourced in compliance with relevant usage guidelines, ensuring no violation of privacy. We have taken care to avoid any biases or discriminatory outcomes in our research process. No personally identifiable information was used, and no experiments were conducted that could raise privacy or security concerns. We are committed to maintaining transparency and integrity throughout the research process.

REPRODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our work. The core components of VoxelMem, including 4D voxel representation, voxel-based control, hybrid global–local memory, and streaming generation with interleave-forcing, are described in Sec. 3. Additional details on video–voxel data construction, model variants, baseline adaptation, benchmark construction, and evaluation protocols are provided in the supplementary material. We further report comprehensive quantitative comparisons and ablation studies to validate the individual design choices. We will publicly release the source code, model weights, and evaluation benchmark upon publication.

REFERENCES

Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025a.

Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025b.

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.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, et al. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13–23, 2025.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan Yuille, Leonidas Guibas, et al. Mixture of contexts for long video generation. *arXiv preprint arXiv:2508.21058*, 2025.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024.

Karan Dalal, Daniel Koceja, Jiarui Xu, Yue Zhao, Shihao Han, Ka Chun Cheung, Jan Kautz, Yejin Choi, Yu Sun, and Xiaolong Wang. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17702–17711, 2025.

Zicheng Duan, Jiatong Xia, Zeyu Zhang, Wenbo Zhang, Gengze Zhou, Chenhui Gou, Yefei He, Feng Chen, Xinyu Zhang, and Lingqiao Liu. Liveworld: Simulating out-of-sight dynamics in generative video world models. *arXiv preprint arXiv:2603.07145*, 2026.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, Chen Sun, et al. Motion prompting: Controlling video generation with motion trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1–12, 2025.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2025.

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

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

Junchao Huang, Xinting Hu, Boyao Han, Shaoshuai Shi, Zhuotao Tian, Tianyu He, and Li Jiang. Memory forcing: Spatio-temporal memory for consistent scene generation on minecraft. *arXiv preprint arXiv:2510.03198*, 2025a.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025b.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598*, 2025.

Bingliang Li, Zhenhong Sun, Jiaming Bian, Yuehao Wu, Yifu Wang, Hongdong Li, Yatao Bian, Huadong Mo, and Daoyi Dong. Storyblender: Inter-shot consistent and editable 3d storyboard with spatial-temporal dynamics. In *European Conference on Computer Vision*, pp. 130–147. Springer, 2026.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. *arXiv preprint arXiv:2503.16421*, 2025a.

Runjia Li, Philip Torr, Andrea Vedaldi, and Tomas Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25690–25699, 2025b.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025c.

Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Yuhao Liu, Tengfei Wang, Fang Liu, Zhenwei Wang, and Rynson WH Lau. Shape-for-motion: Precise and consistent video editing with 3d proxy. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Zhiheng Liu, Xueqing Deng, Shoufa Chen, Angtian Wang, Qiushan Guo, Mingfei Han, Zeyue Xue, Mengzhao Chen, Ping Luo, and Linjie Yang. Worldweaver: Generating long-horizon video worlds via rich perception. *Advances in Neural Information Processing Systems*, 38:46268–46293, 2026.

Wan-Duo Kurt Ma, John P Lewis, and W Bastiaan Kleijn. Trailblazer: Trajectory control for diffusion-based video generation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025.

Yihao Meng, Hao Ouyang, Yue Yu, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Hanlin Wang, Yixuan Li, Cheng Chen, Yanhong Zeng, et al. Holocine: Holistic generation of cinematic multi-shot long video narratives. *arXiv preprint arXiv:2510.20822*, 2025.

Jacopo Panerati, Hehui Zheng, SiQi Zhou, James Xu, Amanda Prorok, and Angela P Schoellig. Learning to fly—a gym environment with pybullet physics for reinforcement learning of multi-agent quadcopter control. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7512–7519. IEEE, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

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

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025a.

Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–10, 2025b.

Weijie Wang, Haoyu Zhao, Yifan Yang, Feng Chen, Zeyu Zhang, Yefei He, Zicheng Duan, Donny Y Chen, Yuqing Yang, and Bohan Zhuang. Latent spatial memory for video world models. *arXiv preprint arXiv:2606.09828*, 2026a.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026b.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *Advances in Neural Information Processing Systems*, 38:49371–49393, 2026.

Weijia Wu, Zhuang Li, Yuchao Gu, Rui Zhao, Yefei He, David Junhao Zhang, Mike Zheng Shou, Yan Li, Tingting Gao, and Di Zhang. Draganything: Motion control for anything using entity representation. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2024.

Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Yang Zhao, Shanchuan Lin, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. Videoauteur: Towards long narrative video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19163–19173, 2025.

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20406–20417, 2024.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Tooncrafter: Generative cartoon interpolation. *ACM Transactions on Graphics (TOG)*, 43 (6):1–11, 2024.

Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2025a.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2025b.

Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. Direct-a-video: Customized video generation with user-directed camera movement and object motion. *arXiv preprint arXiv:2402.03162*, 2024a.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

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*, 2024b.

Jung Yi, Wooseok Jang, Paul Hyunbin Cho, Jisu Nam, Heeji Yoon, and Seungryong Kim. Deep forcing: Training-free long video generation with deep sink and participative compression. *arXiv preprint arXiv:2512.05081*, 2025.

Jung Yi, Minjae Kim, Paul Hyunbin Cho, Wooseok Jang, Sangdoo Yun, and Seungryong Kim. Worldkv: Efficient world memory with world retrieval and compression. *arXiv preprint arXiv:2605.22718*, 2026.

Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–11, 2025.

Jiwen Yu, Jianxiong Gao, Jianhong Bai, Yiran Qin, Kaiyi Huang, Quande Liu, Xintao Wang, Pengfei Wan, Kun Gai, and Xihui Liu. Memlearner: Learning to query context memory for video world models. *arXiv preprint arXiv:2606.31734*, 2026a.

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.

Wei Yu, Runjia Qian, Yumeng Li, Liquan Wang, Songheng Yin, Dennis Anthony, Yang Ye, Yidi Li, Weiwei Wan, Animesh Garg, et al. Mosaicmem: Hybrid spatial memory for controllable video world models. *arXiv preprint arXiv:2603.17117*, 2026b.

Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv e-prints*, pp. arXiv–2504, 2025.

Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Pretraining frame preservation in autoregressive video memory compression. *arXiv preprint arXiv:2512.23851*, 2025a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2063–2073, 2025b.

Zhiyuan Zhang, Dongdong Chen, and Jing Liao. I2v3d: Controllable image-to-video generation with 3d guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13360–13371, 2025c.

Zhiyuan Zhang, Can Wang, Dongdong Chen, and Jing Liao. Flextraj: Image-to-video generation with flexible point trajectory control. *arXiv preprint arXiv:2510.08527*, 2025d.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. *arXiv preprint arXiv:2601.05138*, 2026.

Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, and Tong He. Aether: Geometric-aware unified world modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8535–8546, 2025.

A APPENDIX

This supplementary material provides additional details on the video–voxel data construction pipeline, the VM-4D benchmark, baseline adaptation, and evaluation protocols used in the main paper.

A.1 VIDEO–VOXEL DATASET CONSTRUCTION

As described in Sec. 4 of the main paper, we construct paired video–voxel data from both real-world videos and synthetic rendered scenes. The two sources are complementary: real videos provide realistic appearance and motion, while synthetic scenes provide accurate and consistent geometry.

Real-world data. For real-world videos, we first obtain object and region segmentation using SAM2 (Ravi et al., 2024) and estimate long-range 3D point trajectories using SpatialTracker (Xiao et al., 2024). Tracking is particularly important for dynamic regions, as it provides temporally consistent correspondences for objects that move or deform over time.

We convert tracked points into voxel-like primitives through motion-aware clustering within each segmentation region. For tracked points i and j, we define their distance as

$$D_{ij} = \mathrm{Std}_{t\in\Omega_{ij}}\left(\|\mathbf{p}_i^t - \mathbf{p}_j^t\|_2\right) + \lambda\|\mathbf{p}_i^0 - \mathbf{p}_j^0\|_2, \tag{10}$$

where Ω_{ij} denotes the frames in which both points are visible. The first term measures the temporal variation of their relative distance and therefore favors points that move coherently, while the second term encourages spatially nearby points to be grouped together.

Clustering is performed within segmentation regions to avoid merging points from different object instances. A granularity factor controls the clustering resolution, allowing the same pipeline to produce primitives ranging from coarse instance-level regions to fine point-level structures. Each resulting cluster is converted into a voxel primitive according to its spatial center and extent.

Synthetic data. For synthetic training data, we collect more than a dozen Blender scenes and hundreds of animated characters. Static meshes are voxelized using an axis-aligned bounding-box (AABB) grid with BVH-based occupancy queries. For dynamic objects, we construct persistent vertex groups whose membership remains fixed throughout the animation, producing temporally coherent voxel trajectories as the underlying mesh deforms.

We render each scene under randomized HDR environments and smooth camera trajectories, and export the RGB video, voxel trajectories, and camera parameters. The resulting synthetic data provide accurate geometry that complements the noisier geometry estimated from real-world videos.

A.2 MODEL VARIANTS

We train bidirectional Wan2.2 (Wan et al., 2025) models at both 5B and 14B scales. The 14B model is included to enable comparison with baselines using similarly sized backbones. We further distill the 5B model into a causal streaming generator following the procedure described in the main paper. All streaming results use the distilled 5B model, while all ablation studies use the bidirectional 5B model. We additionally report the bidirectional 5B variant in the main comparison to separate the effects of architecture and model scale.

A.3 BASELINE ADAPTATION

We compare against two groups of methods: geometry-aware control methods and long-video generation methods. FlexTraj (Zhang et al., 2025d), Spmem (Wu et al., 2026), and Matrix-Game3 (Wang et al., 2026b) use 5B-scale backbones, while VerseCrafter (Zheng et al., 2026) and LiveWorld (Duan et al., 2026) use 14B-scale backbones.

Geometry-aware control methods. We compare with FlexTraj (Zhang et al., 2025d) and VerseCrafter (Zheng et al., 2026). These methods are originally designed for I2V generation and do not explicitly maintain long-term memory. To extend them to long-video generation, we generate videos chunk by chunk and use the last frame of each chunk as the input image for the next chunk.

Long-video generation methods. We compare with Matrix-Game3 (Wang et al., 2026b), Spmem (Wu et al., 2026), and LiveWorld (Duan et al., 2026).

Matrix-Game3 uses keyboard actions rather than explicit camera trajectories. We therefore convert our camera poses into keyboard-style control signals following Yume (Mao et al., 2025).

Spmem and LiveWorld rely on point-cloud-based spatial representations. To provide comparable geometric information, we convert our voxel fields into dense point clouds using multi-frame ray casting from the voxel geometry and use the resulting point clouds as their spatial inputs.

A.4 VM-4D Revisit Benchmark

We construct VM-4D to evaluate long-range consistency under camera revisits and dynamic scene evolution. The benchmark contains 360 scenes, each with 401 frames. Each scene provides a high-quality rerendered input image, a camera trajectory, and corresponding static and dynamic voxel controls. We design 16 camera trajectories with different moving-out-and-return patterns to cover diverse revisit behaviors. We consider two revisit settings.

Same-pose revisit. This setting contains 177 scenes. The camera follows an $A \rightarrow B \rightarrow A \rightarrow B \rightarrow A$ trajectory, repeatedly leaving and revisiting previously observed regions. It primarily evaluates whether a model can recover scene appearance and geometry after content moves out of view and later reappears.

Different-pose revisit. This setting contains 183 scenes. Instead of returning to exactly the same camera pose, the camera revisits the same scene or object region from different viewpoints. This setting evaluates consistency under more challenging viewpoint changes.

A.5 VM-4D Evaluation Metrics

Because ground-truth future videos are generally unavailable for out-of-sight dynamic content, we follow a reference-based evaluation protocol (Duan et al., 2026).

Revisit consistency. For same-pose revisits, we evaluate static scene regions using PSNR, SSIM, and LPIPS (Zhang et al., 2018) at two revisit positions, denoted as RetA and RetB.

RetA compares the first and final observations of region A, corresponding to content visible in the initial view. RetB compares the first and second observations of region B, corresponding to a region first revealed after the camera moves away from the initial view. These measurements evaluate whether previously observed content can be recovered consistently after leaving the field of view.

Geometric control. We evaluate geometric consistency using depth RMSE (D-RMSE) following (Eigen et al., 2014). Lower D-RMSE indicates better agreement between the generated geometry and the target geometric control. The depth is estimated by Depth-anything-3 (Lin et al., 2025).

Video consistency. We compute CLIP image similarity (CLIP-I (Radford et al., 2021)) between consecutive frames, separately for foreground and background regions. This evaluates temporal appearance consistency for dynamic objects and static scene content.

A.6 VBench-I2V Evaluation

In addition to VM-4D, we evaluate general image-to-video generation quality on 450 videos from SpatialViD (Wang et al., 2025a), filtered following VerseCrafter (Zheng et al., 2026). We use the VBench-I2V evaluation protocol and report eight dimensions covering visual quality, motion, temporal consistency, and image-conditioned consistency. The overall score reported in the main paper is the arithmetic mean of these eight dimensions.

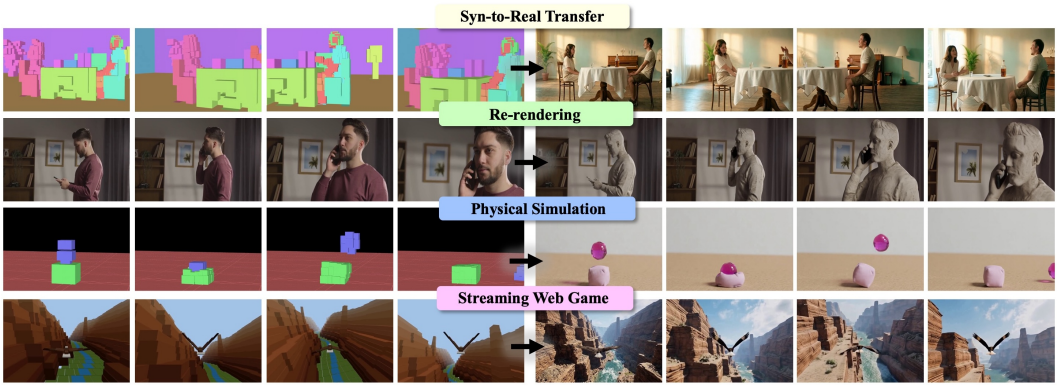


Figure 8: More examples, showing consistent video synthesis across voxel-controlled scenarios.

A.7 More Results and Applications

We provide **additional results on a static project webpage** for easier inspection of diverse applications. The longest videos evaluated in our experiments are 30 seconds at 16 FPS, corresponding to around 500 generated frames.

Filmmaking applications. For syn-to-real transfer, we collect Blender scenes and animate foreground characters with Mixamo motions to obtain controllable voxel trajectories and camera paths, including scenes constructed with StoryBlender (Li et al., 2026). For re-rendering, we estimate voxel trajectories directly from source videos. For both settings, the initial frame is first re-rendered with Nano Banana 2, and VoxelMem then generates the video with the corresponding voxel controls.

Physical simulation. For physics-driven generation, we construct synthetic scenes using Pybullet (Panerati et al., 2021) engine, which provides rigid-body physics simulation for dynamic multi-object scenes. We obtain object trajectories (rotation/translation) from the simulator and convert the resulting time-varying geometry into voxel controls. VoxelMem then generates realistic videos conditioned on these simulated dynamics, enabling physically plausible video generation.

Streaming web-game generation. For the streaming application, we use Gemini 3.0 Pro to generate Three.js code for constructing voxel-style web games, including the 3D scene, game logic, and interactive controls. User interactions determine the future scene and camera evolution, which are converted into voxel controls and passed to the VoxelMem for streaming generation.