AnchorWeave: World-Consistent Video Generation with Retrieved Local Spatial Memories

Zun Wang[1], Han Lin[1], Jaehong Yoon[3], Jaemin Cho[2,4], Yue Zhang[1], and Mohit Bansal[1]

[1] University of North Carolina at Chapel Hill, USA
[2] Johns Hopkins University, USA
[3] Nanyang Technological University, Singapore
[4] Allen Institute for AI, USA
https://zunwang1.github.io/AnchorWeave

Abstract. Maintaining spatial world consistency over long horizons remains a central challenge for camera-controllable video generation. Existing memory-based approaches often condition generation on globally reconstructed 3D scenes by rendering anchor videos from the reconstructed geometry in the history. However, reconstructing a global 3D scene from multiple views inevitably introduces cross-view misalignment, as pose and depth estimation errors cause the same surfaces to be reconstructed at slightly different 3D locations across views. When fused, these inconsistencies accumulate into noisy geometry that contaminates the conditioning signals and degrades generation quality. We introduce **AnchorWeave**, a memory-augmented video generation framework that replaces a single misaligned global memory with multiple clean local geometric memories and learns to reconcile their cross-view inconsistencies. To this end, AnchorWeave performs coverage-driven local memory retrieval aligned with the target trajectory and integrates the selected local memories through a multi-anchor weaving controller during generation. Extensive experiments demonstrate that AnchorWeave significantly improves long-term scene consistency while maintaining strong visual quality, with ablation and analysis studies further validating the effectiveness of local geometric conditioning, multi-anchor control, and coverage-driven retrieval. We provide more qualitative demonstration videos in the supplementary material via an anonymized local webpage.

Keywords: Long-Horizon Video Generation · Memory-Augmented Generation · Spatial World Consistency · World Modeling

1 Introduction

Building video world models that can consistently generate and revisit the same environments over long horizons under user control (such as keyboard directions [5, 8, 68], camera viewpoints [14, 29, 87], or human/robot actions [27] re-

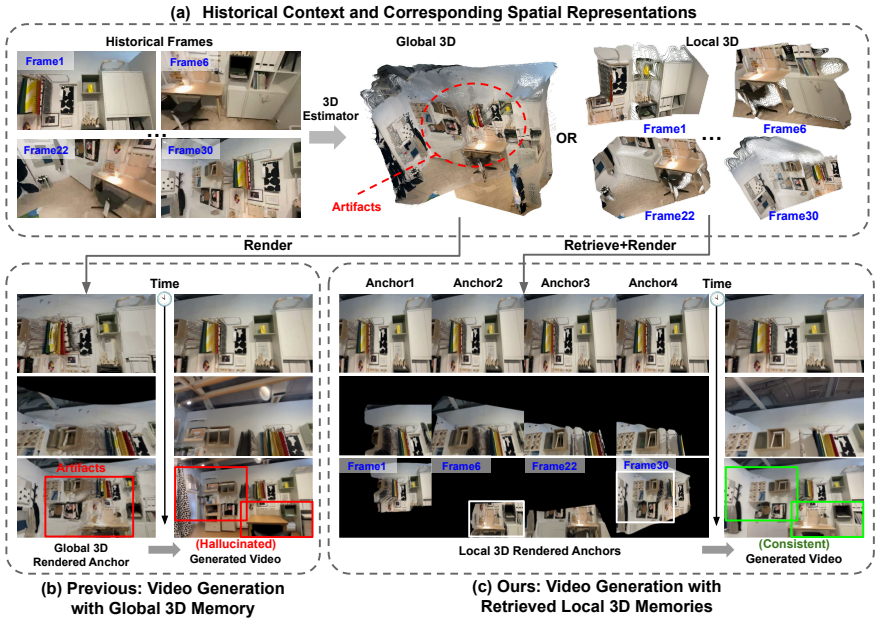

Fig. 1: Global 3D reconstruction accumulates cross-view misalignment, introducing artifacts in the reconstructed geometry ((a), middle), which propagate to the generated video as hallucinations ((b), red boxes). In contrast, per-frame local geometry inherently avoids cross-view misalignment and therefore remains clean ((a), right). Conditioning on multiple retrieved local geometric anchors, AnchorWeave maintains strong spatial consistency with the historical frames ((c), white and green boxes).

mains a central challenge in generative modeling. Recent image-to-video diffusion and Transformer-based models [28,45,83] have significantly improved short-term coherence and controllability, enabling clip-by-clip autoregressive generation [12, 40, 114] that follows camera trajectories. However, such models usually lack persistent memory mechanisms, causing them to lose consistency with previously generated content when revisiting earlier regions of the environment [108].

To improve long-horizon world-consistent video generation, recent works [48, 92, 118] introduce explicit memory mechanisms that reconstruct a global 3D scene from historical video clips. Specifically, they estimate camera poses and per-frame geometry (e.g., depth maps), and fuse them into a unified 3D representation such as a global point cloud. Generation is then conditioned on anchor videos (i.e., videos rendered from the reconstructed 3D scene under target camera trajectories) to enable spatial consistency. However, even state-of-the-art 3D reconstruction models struggle to maintain accurate global geometry under dense, long-horizon views, making precise cross-view alignment difficult. Small pose or depth estimation errors cause the same surfaces to be reconstructed at slightly different 3D locations across views (Figure 1(a) middle artifacts); when fused, these inconsistencies introduce structured noise into rendered an-

chor videos, leading to ghosting or drift artifacts that degrade generation quality (e.g., inconsistent or hallucinated content in Figure 1(b)).

We propose **AnchorWeave**, a 3D-informed, memory-aware video generation framework that avoids relying on a globally fused scene representation. Our key insight is that cross-view misalignment artifacts can be mitigated by replacing a single global 3D memory with per-frame local point cloud memories, as illustrated in Figure 1(c). Since local point clouds do not accumulate ghosting or drift from multi-view fusion, they provide cleaner geometric signals for rendering-based conditioning. While these rendered anchors may still exhibit residual misalignment, we treat their reconciliation during generation as a learnable problem and resolve it through a learned multi-anchor weaving module. Specifically, given a sequence of previously generated or observed frames, AnchorWeave estimates and maintains per-frame local geometry as spatial memory for long-horizon generation. To synthesize new content under a target camera trajectory, AnchorWeave introduces a coverage-driven memory retrieval formulation that iteratively selects the local memory that maximizes additional visibility coverage along the target trajectory. The selected memories are rendered as multiple anchor videos for conditioning. A multi-anchor weaving controller with shared cross-anchor attention and pose-guided fusion then *weaves* these anchors into a unified control signal, enabling effective multi-anchor conditioning while mitigating residual geometric inconsistencies. The newly generated frames are updated to the history for subsequent generation, enabling AnchorWeave to iteratively extend the video through the update–retrieve–generate loop.

We compare AnchorWeave with multiple baselines spanning different memory conditioning paradigms, including global point cloud and implicit cross-view conditioning, on RealEstate10K and DL3DV under a history-conditioned video generation setting. AnchorWeave significantly improves visual quality and long-horizon scene consistency, while generalizing well to open-domain images and scenes. Ablation studies further validate the contribution of each component, including local geometric memory, coverage-driven retrieval, and the geometry-conditioned multi-anchor controller, supporting the effectiveness of conditioning on local geometry and learning to handle misalignment.

2 Related Work

Camera-controllable video generation. Recent advances in video generation with camera controllability span text-to-video (T2V), image-to-video (I2V), and video-to-video (V2V) settings. Early approaches primarily inject explicit camera parameterizations (e.g., Plücker embeddings or pose tokens) into video diffusion models for conditioning [1, 2, 30, 31, 37, 49, 51, 52, 54, 55, 78, 86, 87, 89, 99, 111, 112, 119, 121]. To improve geometric faithfulness, subsequent works introduce structured 3D guidance, such as conditioning on rendered point clouds or anchor videos [6, 10, 36, 39, 47, 58, 65, 70, 71, 75, 101, 110, 113, 115, 117, 120, 124]. Another line of research formulates camera control through trajectory tracking or motion encoding as intermediate guidance [9, 20, 21, 25, 44, 62, 90, 96]. In the V2V regime,

some methods rely on test-time optimization or per-scene fine-tuning [67,76,105, 113], while others construct large-scale paired data from simulators such as Unreal Engine5, Kubric, or Animated Objaverse [3,4,19,22,23,42,81,82,84,91,106]. Recent Approaches also incorporate structured 3D priors for controllable V2V generation [7, 16, 38, 61, 66, 98, 102, 109]. Despite substantial progress, existing camera-controllable video models are typically memoryless: they condition on per-frame parameters, rendered anchors, or short local context, but lack an explicit, robust memory mechanism to maintain long-horizon spatial consistency under complex camera motions. In contrast, we focus on equipping camera control models with a geometry-aware and retrieval-based memory design, enabling robust long-horizon consistency and controllability beyond single-shot or short-context conditioning.

Memory-augmented video generation. Recent video world models extend long-horizon consistency by incorporating memory beyond the limited context window of current generators. One line of work adopts retrieval-based mechanisms, treating past frames as external memory and retrieving relevant history for future synthesis [17,24,26,50,95,97,108]. Another line preserves spatial structure through explicit 3D memory constructed from historical inputs, such as maintaining global point clouds to support revisit consistency and camera control [48, 48, 53, 60, 90, 92, 107, 118, 122]. A third direction leverages recurrent or state-space models to enhance long-horizon temporal modeling [18, 35, 46, 69, 74, 116]. Several works also improve consistency or long-term quality by redesigning attention or latent conditioning in autoregressive or diffusion-based models [11, 13, 32, 34, 41, 59, 63, 64, 73, 77, 79, 80, 85, 93, 94, 100, 104, 114]. Among them, the closest are 3D-based approaches including SPMem [92], Spatia [118], MagicWorld [48] that rely on global 3D point cloud as memory. Instead of noisy global reconstruction, we represent memory as local geometric anchors and perform multi-anchor joint conditioning for more stable long-horizon consistency.

3 Methodology: AnchorWeave

AnchorWeave is a world-consistent video generation framework that leverages retrieved local spatial memories for long-horizon, camera-controlled generation. It builds upon a standard video diffusion backbone (Section 3.1) augmented with explicit spatial memory. Instead of fusing observations into a single global 3D memory (*i.e.,* global point cloud), AnchorWeave maintains a collection of per-frame local geometric memories, each represented as a local point cloud with an associated camera pose (Section 3.2). Given this memory bank, relevant memories are selected for a target camera trajectory using a coverage-driven, 3D-aware retrieval strategy (Section 3.3) and rendered as anchor videos. These anchor videos are integrated into the diffusion process through a Multi-anchor Weaving Controller, which jointly reasons over multiple anchors via shared attention and pose-guided fusion (Section 3.4). Finally, following [92] that condition on global memory, AnchorWeave operates in an iterative update–retrieve–generate loop,

where newly generated frames are converted into local geometric memories to support long-horizon and consistent generation (Section 3.5).

3.1 Preliminary: Video Diffusion Models

We use standard DiT-based latent diffusion models (LDMs) [72] as backbone. Specifically, given an RGB video with F frames, $\boldsymbol{x} \in \mathbb{R}^{F\times 3\times H\times W}$, a pretrained video autoencoder first compresses the input into a lower-dimensional latent tensor: $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x}) \in \mathbb{R}^{F\times C'\times H'\times W'}$, where H' and W' denote the spatially downsampled resolution. To enable generative modeling, noise is injected into the latent representation through a predefined scheduling strategy, such as DDPM [33] or Flow Matching [57], producing a noisy latent $\boldsymbol{z}_t$ at timestep t. A conditional diffusion network $\boldsymbol{\mathcal{F}_\theta}(\boldsymbol{z}_t, t, \boldsymbol{c}_{\text{text/img}})$ is then trained to recover the underlying clean signal by reversing this corruption process, where the model is conditioned on the diffusion timestep and auxiliary inputs, including textual descriptions $\boldsymbol{c}_{\text{text}}$ for text-to-video generation, and extra camera intrinsics and anchor videos $\boldsymbol{c}_{\text{cam}}$ for video generation with camera control. Training is performed by minimizing the following objective: $\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\boldsymbol{z},\,\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I}),\,t}\,\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, t, \boldsymbol{c}_{\text{text+cam}})\|_2^2$ where $\boldsymbol{\epsilon}$ represents the sampled Gaussian perturbation and $\boldsymbol{\epsilon}_\theta$ denotes the noise estimate produced by the model. This objective is used to train all diffusion components.

3.2 Local Geometric Memory Construction

We represent **spatial memory as a set of per-frame local point clouds** rather than a single global 3D model, thereby avoiding the accumulation and propagation of cross-view misalignment artifacts. The construction proceeds as follows. We assume access to a history context consisting of previously observed video frames, which can come from real-world inputs or earlier generated segments. These frames serve as the visual context from which spatial memory is constructed and updated throughout long-horizon generation. Using this history context, we estimate per-frame local geometry and camera pose with a pretrained 3D reconstruction model (e.g., TTT3R [15]). Each frame's memory is represented as a local point cloud together with its camera pose, which is transformed into a shared world coordinate system and stored as an independent spatial memory entry. This update can be performed incrementally only for newly generated frames in practice.

3.3 Coverage-driven Memory Retrieval

With a large pool of history local memories, we aim to retrieve a compact subset that can effectively guide generation for a target camera trajectory. We illustrate the retrieval pipeline in Figure 2 and provide the complete retrieval procedure in the Appendix.

Retrieval pipeline. Given a target camera trajectory of length T for the next video segment, we divide it into T/D temporal chunks, each spanning D

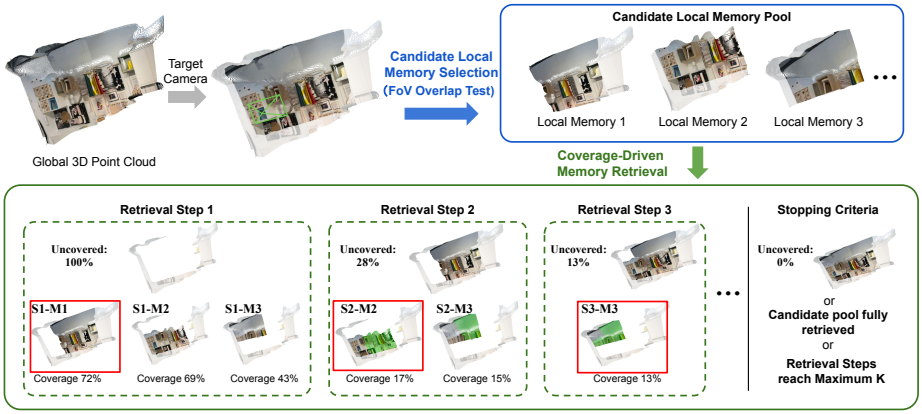

Fig. 2: Coverage-driven memory retrieval pipeline. Given a target camera, we first select local memories whose camera FoVs partially overlap with the target camera view to form a candidate memory pool. At each retrieval step, we greedily select the memory that maximizes the newly covered visible area. Points invisible to the target camera are highlighted in white (For example, in the top row, the white regions in the second point cloud (compared to the first) correspond to areas outside the target camera's field of view as invisible regions). Si–Mj denotes memory j selected at retrieval step i; the red box indicates the retrieved memory. In Si-Mj, regions already covered by previously retrieved memories are marked in green. Only newly covered regions retain their original RGB colors. No green regions appear in S1-Mj since 1st-step's coverage is empty. Retrieval terminates when uncovered region is 0%, the retrieval budget K is exhausted, or the remaining memory pool is empty. For clarity, coverage is computed with a single frame here, while in practice is aggregated over multiple frames per chunk.

frames, and perform memory retrieval independently for each chunk. We formulate memory retrieval as a **visibility coverage maximization** problem: the selected local memories should jointly maximize coverage along the target camera trajectory within each chunk. Specifically, as shown in the upper part of Figure 2, for each chunk, we first construct a candidate local memory pool using a coarse field-of-view (FoV) overlap test to filter out clearly irrelevant local memories. Among the remaining candidates, we iteratively select the local memory that provides the largest new coverage of the currently uncovered region along the chunk trajectory, shown in the bottom part of Figure 2, Retrieval stops when the visible region is fully covered, the candidate pool is exhausted, or the number of retrieved memories reaches the maximum K. This greedy strategy yields a compact and complementary set of local memories per chunk, avoiding redundant memories that offer little additional guidance.

Anchor-video construction and pose computation. For each chunk, we retrieve up to K local memories. For each retrieved memory, we render anchor views along the target camera trajectory within the chunk, producing anchor clips of length D frames. If fewer than K memories are retrieved, we pad the remaining anchor slots with fully-invisible anchor clips to keep a fixed number of K anchors per chunk.

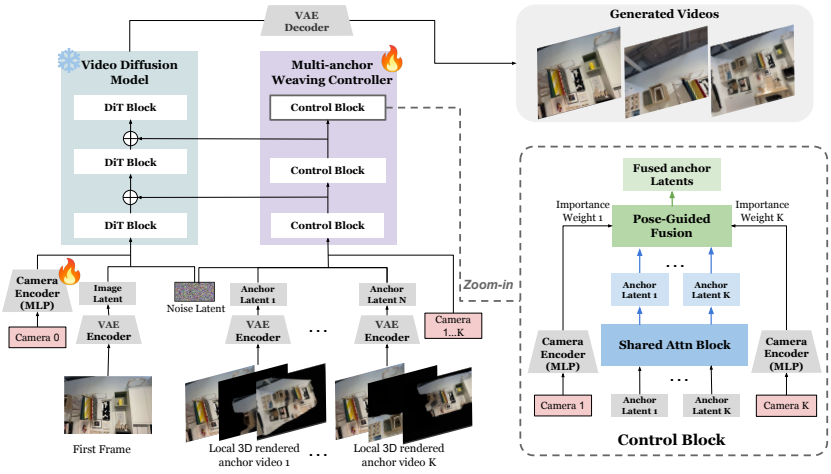

Fig. 3: Architecture of multi-anchor weaving controller. Anchors are encoded and jointly processed by a shared attention block, followed by camera-pose-guided fusion to produce a unified control signal injected into the backbone model. Camera 1 to K represent the retrieved-to-target camera poses for the 1 to K anchor videos, where each denotes the relative pose between the camera associated with a retrieved local point cloud and the target camera, measuring their viewpoint proximity. Camera 0 is the relative target camera trajectory.

For the rendered anchors, we compute the retrieved-to-target relative camera pose sequence, defined as the rigid transformations between the camera pose at which the local memory was captured and the target camera poses within the chunk. Anchor clips corresponding to the same retrieval index are then concatenated across chunks, resulting in K anchor videos of length T with temporally aligned retrieved-to-target pose trajectories. These anchor videos and their pose sequences are used as conditioning signals for generation.

3.4 Generation with Multi-Anchor Weaving Controller

Given K retrieved anchor videos, we condition generation on their latents together with retrieved-to-target relative camera poses, and the target camera pose, enabling the model to *weave* multiple anchor conditions into a coherent and spatially consistent output video.

Overall architecture. As shown in Figure 3, each retrieved anchor video is first encoded into latent features using the same 3D VAE as the backbone diffusion model. These anchor latents are then fed into a multi-anchor controller, composed of a stack of DiT-based ControlNet blocks that operate alongside the backbone model. At each denoising layer, we enable joint multi-anchor modeling through two key designs: (1) a shared multi-anchor attention module that jointly processes all anchor features, and (2) a camera-pose-guided fusion module that aggregates anchor features into a single control signal. The resulting fused control

feature is injected into the corresponding layer of the video diffusion backbone to provide geometric guidance. Details of these designs are presented below.

Shared multi-anchor attention. Instead of processing anchors independently, we jointly process the K anchor latents using a shared-attention block (Figure 3 inside the control block). Concretely, we reshape the K anchor latent sequences with shape $L_a \times C_a$ into a single sequence of shape $1 \times (K \cdot L_a) \times C_a$ (L_a denotes the number of latent tokens per anchor and C_a is the hidden dimension), and apply joint attention over the concatenated tokens. This joint attention enables information exchange across anchors, allowing the controller to aggregate complementary geometric evidence while suppressing contradictory or noisy signals, rather than committing to any single anchor rendering.

Pose-aware importance estimation and fusion. Since not all anchors contribute equally to a target view, we condition anchor fusion on the retrieved-to-target relative camera poses (Figure 3 right, camera 1 to K). At each control layer, these camera poses are encoded by a learnable MLP to produce per-anchor importance weights, which reflect the geometric proximity between the target view and the cameras associated with the retrieved local point clouds used to render the anchors. Using these importance weights, we perform a weighted sum over the anchor-conditioned features to get a single fused control feature, which is injected into the corresponding backbone block to provide geometric guidance.

Target camera pose control. We further condition the model on the target camera trajectory (denoted as camera 0 in Figure 3) to handle cases where geometric memories provide limited guidance, such as early generation stages with rapid camera motion that moves the target view beyond all anchor visibility. The target camera pose is encoded by a trainable camera encoder to produce camera embeddings, which are injected into the backbone DiT blocks as explicit camera-motion control signals. This complementary signal improves camera following under long-horizon generation.

Overall, the multi-anchor weaving controller explicitly conditions generation on geometry through joint attention, pose-conditioned fusion, and target camera pose control, transforming imperfect anchors into a coherent conditioning signal.

3.5 Long-horizon Generation with Updated Memories

Each iteration of the components described in Sections 3.2-3.4 produces one video segment. Repeating this process enables long-horizon video generation via an *update–retrieve–generate* loop in AnchorWeave. **(1) Update (Section 3.2):** Starting from an initial frame or short segment, we estimate per-frame local geometry and initialize the spatial memory bank. **(2) Retrieve (Section 3.3):** Given a target camera trajectory, we retrieve relevant local memories using our coverage-driven memory retrieval and condition the model on both the target pose and retrieved anchors. **(3) Generate (Section 3.4):** The backbone model generates the next video segment under the control of the multi-anchor geometry-aware controller. The newly generated frames are then added to the memory bank, enabling memory-augmented generation over long trajectories.

4 Experiments

4.1 Experiment Setup

Implementation details. We train AnchorWeave on 10K videos sampled from RealEstate10K [123] and DL3DV [56]. For each video, per-frame local geometry is estimated using TTT3R [15] and stored as local point-cloud memories. During training, we use chunk length $D = 8$, retrieving memories every 8 frames with at most $K = 4$ point clouds for conditioning. To improve robustness, we randomly sample from the candidate memory pool as condition augmentation. For memory retrieval, the first rendered frame is always selected as the initial anchor, following the standard anchor-video convention [71, 109, 113]. We apply AnchorWeave to two DiT-based backbones: CogVideoX-I2V-5B [103] and Wan2.2-TI2V-5B [83]. The multi-anchor controller is injected into the first third of backbone layers and applied during the first 90% of denoising steps. We use random frame masking on anchor videos to improve robustness to imperfect geometric guidance. The model is trained with Adam (lr=2×10^{-4}) for 10K steps with batch size 8. Additional details are provided in the Appendix.

Evaluation setup and dataset. We evaluate the model's ability to condition on historical context under a *partial-revisit* setting. We sample videos with large camera motion (with substantial viewpoint changes between frames) from RealEstate10K [123] and DL3DV [56], collecting a total of 500 videos for evaluation. For each video, we sample 70 frames, uniformly select 49 frames as the target segment, and use the remaining 21 frames as historical context available for retrieval. This enables direct quantitative and qualitative comparison against ground-truth (GT) frames. In addition, we include long-horizon qualitative examples using keyboard-style camera actions starting from a single open-domain initial frame, assessing long-term controllability and scene consistency.

Evaluation metrics. Following [92], we use PSNR and SSIM to measure reconstruction fidelity, computed on predicted and GT frames. For perceptual quality, the VBench protocol [43] is adopted, including Subject/Background Consistency, Motion Smoothness, Temporal Flickering, Aesthetic /Imaging Quality.

4.2 Comparison with Prior Methods

Baselines. We compare AnchorWeave with state-of-the-art camera-controllable, memory-augmented video generation methods, including Gen3C [71], SEVA [121], ViewCrafter [110], TrajCrafter [109], EPiC [88], Context-as-Memory [108], and SPMem [92]. Gen3C, TrajCrafter, ViewCrafter, and EPiC are designed to condition generation on a single anchor video. To provide strong baselines, we construct the anchor video for these methods using our retrieval strategy with K=1, where the retrieved local memory corresponds to the local point cloud that maximizes spatial coverage for each chunk. This allows these baselines to condition on the most informative retrieved local memory, in the same spirit as our method. We additionally include a single-anchor variant of AnchorWeave ($K = 1$, one retrieval per chunk) to directly compare with these baselines under identical

Table 1: Quantitative results on RealEstate10K and DL3DV under partial-revisit evaluation. The best numbers are in **bold** and the second best numbers are underlined. The total quality score is the average of all quality dimensions. † denotes our reimplementation based on CogVideoX with the same data as ours (due to not open-sourced). K denotes the maximum retrieval per chunk.

Method	Quality Score (↑)							Consistency Score (↑)	
	Total	Subject Consist	Bg Consist	Motion Smooth	Temporal Flicker	Aesthetic Quality	Imaging Quality	PSNR	SSIM
ViewCrafter [110]	78.58	85.50	90.11	90.91	91.36	46.71	**66.92**	16.17	0.5240
Gen3C [71]	77.11	84.22	89.08	93.74	90.43	45.25	59.92	18.23	0.5926
TrajCrafter [109]	76.34	83.56	88.94	91.43	88.06	45.85	60.19	16.96	0.5767
EPiC [88]	77.77	85.22	90.08	94.74	90.43	45.25	60.92	17.42	0.5704
Context-as-Memory† [108]	78.07	85.08	89.94	92.36	90.21	46.97	63.88	17.91	0.5884
SPMem† [92]	76.85	84.52	89.31	91.84	89.67	45.61	60.12	17.25	0.5710
SEVA [121]	79.66	87.64	92.31	94.31	91.00	48.90	63.78	**21.13**	0.6711
AnchorWeave (CogVideoX-5B, K=1)	80.07	87.70	92.51	96.33	93.05	47.88	62.94	19.01	0.6145
AnchorWeave (CogVideoX-5B)	80.30	87.89	92.34	96.21	93.90	47.87	63.61	20.96	0.6727
AnchorWeave (Wan2.2-5B, K=1)	80.76	88.05	**92.78**	**96.58**	93.32	**49.40**	64.45	19.60	0.6180
AnchorWeave (Wan2.2-5B)	**80.98**	**88.25**	92.60	96.45	**94.15**	49.35	65.10	21.04	**0.6739**

conditioning settings. Context-as-Memory and SEVA directly take multi-view history as conditions for memory consistency, while SPMem conditions generation on global point cloud renderings together with selected key historical frames. Note that, as Context-as-Memory and SPMem are not open-sourced, we reimplement them on the CogVideoX backbone using the same training data as our approach for fair comparison. Additional implementation details are provided in the Appendix.

Quantitative comparison. Table 1 shows that AnchorWeave achieves the best overall performance across both reconstruction and perceptual quality metrics. While AnchorWeave and SEVA [121] obtain comparable consistency scores, AnchorWeave exhibits significantly better temporal quality, including smoother motion and reduced flickering, benefiting from preserving the pretrained video backbone's motion prior. Compared to single-anchor baselines, our single-anchor variant ($K = 1$) consistently achieves higher quality and consistency, indicating that our design extracts more fine-grained geometric details from anchor conditioning and corrects residual errors. Our method also consistently surpasses memory-based baselines, including Context-as-Memory and SPMem. Moreover, the full AnchorWeave further improves over its $K = 1$ variant, demonstrating the effectiveness of multiple retrieval to improve consistency. Notably, AnchorWeave already outperforms all baselines when built on the CogVideoX backbone, and replacing CogVideoX with Wan2.2 leads to further performance gains. We also evaluate camera controllability compared to baselines in the Appendix.

Qualitative comparison. Figure 4 shows qualitative comparisons under the partial-revisit setting. We present one representative frame for each case. Additional qualitative comparisons under free-form, keyboard-controlled camera trajectories with historical context are provided in the Appendix. Overall, AnchorWeave maintains more consistent scene structure when revisiting, while baseline methods exhibit visible artifacts or misaligned details. Gen3C, EPiC, and TrajCrafter often produce frames with degraded visual quality, including noticeable blurring or structural distortions. ViewCrafter is prone to hallucinated content

Fig. 4: Qualitative comparison with baselines on DL3DV. K=1 means one retrieval per chunk. Baseline methods suffer from spatial drift and inconsistency in details. In contrast, AnchorWeave (ours) maintains consistency for multiple cases. GT and the historical context are shown for reference. For clarity, representative misaligned regions are highlighted with red boxes for strong baselines (e.g., SEVA).

Table 2: Results of conditioning on global or local Memory.

Method	PSNR ↑	SSIM ↑
AnchorWeave w/ global point cloud	16.31	53.45
AnchorWeave w/ local point clouds (ours)	**20.96**	**67.27**

(2nd/3rd example), while SEVA [121] occasionally fails to preserve fine-grained details (see highlighted red-box regions). Context-as-Memory maintains content and scene consistency but lacks precise camera control, often resulting in noticeable deviations from the GT viewpoint. SPMem conditions on rendered global point clouds, which lead to severe hallucinations that remain difficult to mitigate even with additional key-frame conditioning. In contrast, AnchorWeave preserves structural details more reliably through dense, clean local memory conditioning. The single-anchor variant ($K = 1$) sometimes exhibits misalignment due to limited retrieval (see red boxes), whereas the full model achieves stronger consistency by weaving multiple anchors. Overall, AnchorWeave produces results that are visually closer to the ground truth.

Table 3: Effect of number of retrievals. We vary the number of retrievals K per chunk, resulting in at most $6K$ retrieved frames (i.e., 6, 12, and 24 for K=1, 2, 4, respectively). Retrieved frames are not necessarily unique across chunks and duplicates may occur.

Max. # retrieved frames	PSNR ↑	SSIM ↑
6	19.01	0.6145
12	20.01	0.6435
24	20.96	0.6727

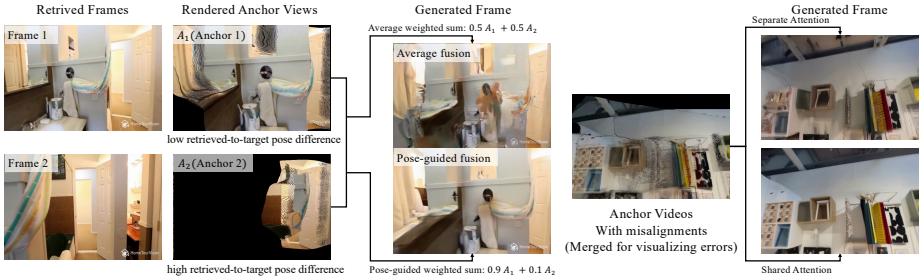

Fig. 5: Ablation study. (a) Pose-guided fusion suppresses misaligned anchors and reduces artifacts compared to simple averaging. (b) Joint attention outperforms separate attention, enabling coherent multi-anchor aggregation.

4.3 Ablations and Analyses

We conduct a series of ablations to analyze the design choices of AnchorWeave, including the use of global vs. local 3D memory, pose-conditioned anchor fusion, joint vs. separate attention in the controller, and the number of retrieved frames. The Appendix further complements these analyses with targeted diagnostics: it visualizes how global point-cloud fusion accumulates drift, stress-tests repeated revisit consistency over 2400 frames, and discusses dynamic-content handling and representative failure modes.

Effect of global vs. local 3D memory. We compare global and local 3D memories as conditioning signals for training the controller. In the global-memory setting, we render a single anchor from the fused global 3D memory, while in the local-memory setting, multiple anchors are rendered from local memories as conditions. The results in Table 2 show that local 3D memory yields improvements in both PSNR and SSIM, confirming the advantage of local geometry for memory conditioning.

Effect of pose-conditioned anchor fusion. As shown in Figure 5(a), simply averaging multiple retrieved anchor views leads to visible artifacts in the generated video when the anchors exhibit different pose deviations from the target view. In contrast, pose-conditioned fusion assigns higher weights to anchors with smaller pose discrepancies, effectively suppressing misaligned anchors and improving visual quality, demonstrating that explicitly accounting for retrieved-to-target pose differences is crucial for robust multi-anchor conditioning.

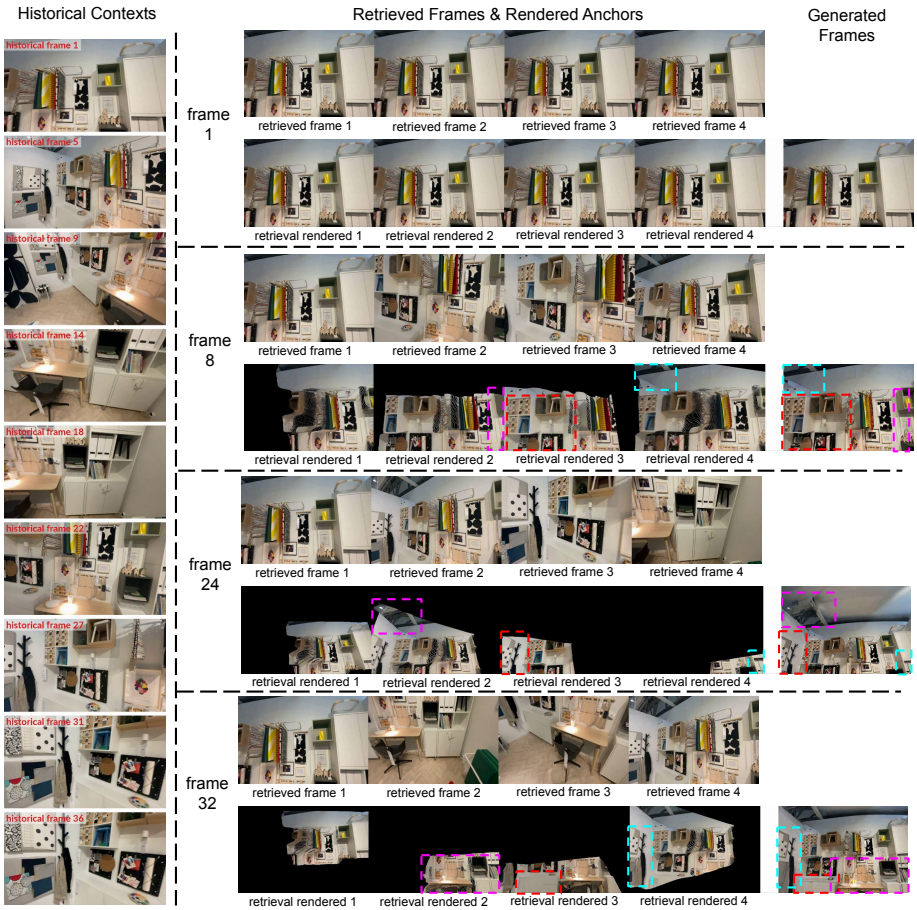

Fig. 6: Visualization of AnchorWeave's retrieval → rendering → generation process. We show representative historical contexts (serving as memory pool), retrieved frames with retrieval rendered anchors, and final generations. For each target frame, all retrievals' anchors provide complementary geometric evidence, and AnchorWeave reconciles them into a coherent output. Same-colored dashed boxes highlight preserved regions.

Effect of joint vs. separate attention in the controller. Figure 5(b) compares the proposed joint-attention-based controller with a baseline that processes each retrieved anchor independently using separate attention blocks. We observe that separate attention leads to blurred or inconsistent structures due to the absence of cross-anchor information exchange. In contrast, joint attention allows the model to aggregate complementary evidence across anchors, resulting in sharper and more coherent geometry in generated frames.

Effect of the number of retrieved frames. Table 3 studies the impact of the number of retrieved frames by varying the number of retrievals K per chunk. Increasing the number of retrieved frames consistently improves PSNR and SSIM, indicating that additional anchors provide complementary spatial evidence for

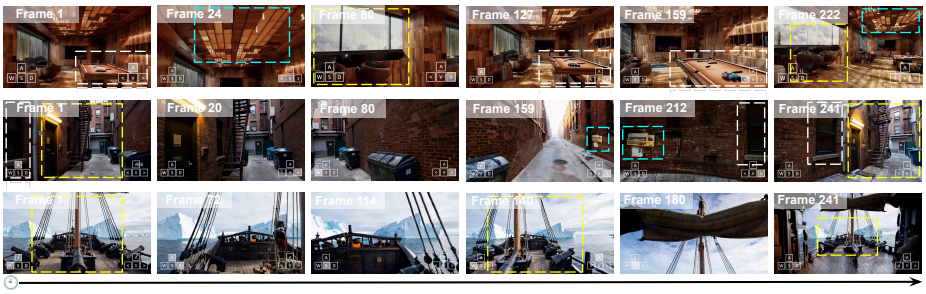

Fig. 7: Long-horizon world exploration examples with keyboard-controlled camera actions on open-domain images. AnchorWeave preserves consistent object geometry and appearance across frames (highlighted by dashed boxes of the same color).

Table 4: Efficiency and performance–cost trade-off. Runtime is measured on an RTX PRO 6000 GPU with Wan2.2-5B at 81×480×832.

Component / setting	Time (s)	GPU (GB)
TTT3R per-frame memory	0.06	–
Retrieval + rendering (ours)	avg. 15	–
Retrieval + rendering (SPMem)	avg. 13	–
Video generation (SPMem)	221	34
Video generation (ours, $K = 4$)	275	39
$K = 1$ (PSNR 19.0)	221	34
$K = 2$ (PSNR 20.0)	238	35
$K = 4$ (PSNR 21.0)	275	39

generation. This trend is also reflected in the comparisons in Figure 4, where more anchors leads to reduced misalignment and improved consistency.

Efficiency and cost. Table 4 summarizes runtime on an RTX PRO 6000 GPU with Wan2.2-5B at 81×480×832 resolution. AnchorWeave adds moderate generation overhead over SPMem due to multi-anchor control (275 vs. 221s; 39 vs. 34GB at $K = 4$), but this comes with a large quality gain over SPMem in Table 1. Increasing K gives a predictable trade-off: moving from $K = 1$ to $K = 4$ improves PSNR by roughly 2dB while adding 54s and 5GB. A full history-length scaling analysis is provided in the Appendix, showing that local-memory retrieval remains more stable as history grows, whereas SPMem's accumulated global-point-cloud rendering cost grows faster.

4.4 AnchorWeave Process Visualization

We visualize the AnchorWeave pipeline (retrieve → render → generate) in Figure 6. For each target frame, multiple coverage-relevant anchors are retrieved and rendered under the target view. Following our implementation details (Section 4.1), the first rendered frame is always selected as the initial anchor, serving as a stable geometric reference. However, under viewpoint changes, it cannot cover all regions previously observed in the historical contexts. Subsequent retrievals (anchors 2–4), selected via coverage-driven retrieval, complement the initial anchor by providing geometric evidence for the regions where the first

frame doesn't cover. Although individual anchors may exhibit misalignment, the joint multi-anchor controller weaves these complementary signals during generation, merging evidence from different anchors into a coherent frame. As a result, information from regions unseen in the first anchor is consistently integrated, preserving strong spatial consistency, as shown in the highlighted dashed boxes with the same colors. Additional examples are provided in the Appendix.

4.5 Open-Domain Long-Horizon Examples

We present three open-domain long-horizon exploration examples in Figure 7 using AnchorWeave with Wan2.2 backbone. The model generalizes effectively to diverse unseen scenarios, such as a wooden cabin, an alley, and a wooden sailing ship (top, middle, and bottom rows, respectively). While the visual backbone is limited to 81-frame video, AnchorWeave enables generation beyond this limit by conditioning each new segment on accumulated memory. As a result, spatial consistency is preserved across multiple segments (e.g. in the cabin scene, regions observed at frame 1 and in later frames (24 and 80), such as the ceiling and stools, remain consistent in the final frame (frame 222); in the alley scene, 360° panoramic consistency is preserved, with strong alignment between the initial and terminal frames). More long-horizon world exploration examples are provided in the Appendix, highlighting long-term consistency, 360° panoramic scene generation, and generalization to third-person gaming scenarios.

5 Conclusion

We present AnchorWeave, a geometry-informed, memory-aware video generation framework that improves long-horizon generation by replacing brittle global 3D reconstruction with multiple clean, view-aligned local geometric memories. Through coverage-driven retrieval and adaptive multi-anchor fusion, AnchorWeave provides robust spatial guidance and significantly improves visual quality, camera controllability, and long-term consistency across diverse scenes and tasks. Our results show that conditioning on local geometric memories and learning to reconcile their inconsistencies offers a reliable alternative to global geometric fusion for scalable long-horizon video generation.

Acknowledgments

This work was supported by DARPA ECOLE Program No. HR00112390060, NSF-AI Engage Institute DRL-2112635, DARPA Machine Commonsense (MCS) Grant N66001-19-2-4031, ARO Award W911NF2110220, ONR Grant N00014-23-1-2356, Accelerate Foundation Models Research program, and a Bloomberg Data Science PhD Fellowship. The views contained in this article are those of the authors and not of the funding agency.

References

1. Bahmani, S., Skorokhodov, I., Qian, G., Siarohin, A., Menapace, W., Tagliasacchi, A., Lindell, D.B., Tulyakov, S.: Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. arXiv preprint arXiv:2411.18673 (2024)
2. Bahmani, S., Skorokhodov, I., Siarohin, A., Menapace, W., Qian, G., Vasilkovsky, M., Lee, H.Y., Wang, C., Zou, J., Tagliasacchi, A., et al.: Vd3d: Taming large video diffusion transformers for 3d camera control. arXiv preprint arXiv:2407.12781 (2024)
3. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025)
4. Bai, J., Xia, M., Wang, X., Yuan, Z., Fu, X., Liu, Z., Hu, H., Wan, P., Zhang, D.: Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. Proc. ICLR (2025)
5. Ball, P.J., Bauer, J., Belletti, F., Brownfield, B., Ephrat, A., Fruchter, S., Gupta, A., Holsheimer, K., Holynski, A., Hron, J., Kaplanis, C., Limont, M., McGill, M., Oliveira, Y., Parker-Holder, J., Perbet, F., Scully, G., Shar, J., Spencer, S., Tov, O., Villegas, R., Wang, E., Yung, J., Baetu, C., Berbel, J., Bridson, D., Bruce, J., Buttimore, G., Chakera, S., Chandra, B., Collins, P., Cullum, A., Damoc, B., Dasagi, V., Gazeau, M., Gbadamosi, C., Han, W., Hirst, E., Kachra, A., Kerley, L., Kjems, K., Knoepfel, E., Koriakin, V., Lo, J., Lu, C., Mehring, Z., Moufarek, A., Nandwani, H., Oliveira, V., Pardo, F., Park, J., Pierson, A., Poole, B., Ran, H., Salimans, T., Sanchez, M., Saprykin, I., Shen, A., Sidhwani, S., Smith, D., Stanton, J., Tomlinson, H., Vijaykumar, D., Wang, L., Wingfield, P., Wong, N., Xu, K., Yew, C., Young, N., Zubov, V., Eck, D., Erhan, D., Kavukcuoglu, K., Hassabis, D., Gharamani, Z., Hadsell, R., van den Oord, A., Mosseri, I., Bolton, A., Singh, S., Rocktäschel, T.: Genie 3: A new frontier for world models. Google DeepMind (2025), `https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/`
6. Bernal-Berdun, E., Serrano, A., Masia, B., Gadelha, M., Hold-Geoffroy, Y., Sun, X., Gutierrez, D.: Precisecam: Precise camera control for text-to-image generation. arXiv preprint arXiv:2501.12910 (2025)
7. Bian, W., Huang, Z., Shi, X., Li, Y., Wang, F.Y., Li, H.: Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. arXiv preprint arXiv:2501.02690 (2025)
8. Bruce, J., Dennis, M.D., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. In: Forty-first International Conference on Machine Learning (2024)
9. Burgert, R., Xu, Y., Xian, W., Pilarski, O., Clausen, P., He, M., Ma, L., Deng, Y., Li, L., Mousavi, M., 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)
10. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. arXiv preprint arXiv:2504.14899 (2025)
11. Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. Advances in Neural Information Processing Systems **37**, 24081–24125 (2024)

12. Chen, B., Monsó, D.M., Du, Y., Simchowitz, M., Tedrake, R., Sitzmann, V.: Diffusion forcing: Next-token prediction meets full-sequence diffusion. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2025)
13. Chen, H., Xu, C., Yang, X., Chen, X., Deng, C.: Past-and future-informed kv cache policy with salience estimation in autoregressive video diffusion. arXiv preprint arXiv:2601.21896 (2026)
14. Chen, T., Hu, X., Ding, Z., Jin, C.: Learning world models for interactive video generation. arXiv preprint arXiv:2505.21996 (2025)
15. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: Ttt3r: 3d reconstruction as test-time training. arXiv preprint arXiv:2509.26645 (2025)
16. Chen, Y., Ye, Z., Fang, Z., Chen, X., Zhang, X., Liu, J., Wang, N., Liu, H., Zhang, G.: Postcam: Camera-controllable novel-view video generation with query-shared cross-attention. arXiv preprint arXiv:2511.17185 (2025)
17. Chou, Y.C., Wang, X., Li, Y., Wang, J., Liu, H., Xie, C., Yuille, A., Xiao, J.: Captain safari: A world engine. arXiv preprint arXiv:2511.22815 (2025)
18. Dalal, K., Koceja, D., Xu, J., Zhao, Y., Han, S., Cheung, K.C., Kautz, J., Choi, Y., Sun, Y., Wang, X.: One-minute video generation with test-time training. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17702–17711 (2025)
19. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
20. Feng, W., Liu, J., Tu, P., Qi, T., Sun, M., Ma, T., Zhao, S., Zhou, S., He, Q.: I2vcontrol-camera: Precise video camera control with adjustable motion strength. arXiv preprint arXiv:2411.06525 (2024)
21. Fortier-Chouinard, F., Hold-Geoffroy, Y., Deschaintre, V., Gadelha, M., Lalonde, J.F.: Gimbaldiffusion: Gravity-aware camera control for video generation. arXiv preprint arXiv:2512.09112 (2025)
22. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. In: Proc. NeurIPS (2024)
23. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., et al.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3761 (2022)
24. Gu, Y., Mao, W., Shou, M.Z.: Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325 (2025)
25. Gu, Z., Yan, R., Lu, J., Li, P., Dou, Z., Si, C., Dong, Z., Liu, Q., Lin, C., Liu, Z., et al.: Diffusion as shader: 3d-aware video diffusion for versatile video generation control. arXiv preprint arXiv:2501.03847 (2025)
26. Guo, X., Ding, C., Dou, H., Zhang, X., Tang, W., Wu, W.: Infinitydrive: Breaking time limits in driving world models. arXiv preprint arXiv:2412.01522 (2024)
27. Guo, Y., Shi, L.X., Chen, J., Finn, C.: Ctrl-world: A controllable generative world model for robot manipulation. arXiv preprint arXiv:2510.10125 (2025)
28. HaCohen, Y., Brazowski, B., Chiprut, N., Bitterman, Y., Kvochko, A., Berkowitz, A., Shalem, D., Lifschitz, D., Moshe, D., Porat, E., et al.: Ltx-2: Efficient joint audio-visual foundation model. arXiv preprint arXiv:2601.03233 (2026)
29. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101 (2024)

30. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for video diffusion models. In: The Thirteenth International Conference on Learning Representations (2025)
31. He, H., Yang, C., Lin, S., Xu, Y., Wei, M., Gui, L., Zhao, Q., Wetzstein, G., Jiang, L., Li, H.: Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. arXiv preprint arXiv:2503.10592 (2025)
32. He, X., Peng, C., Liu, Z., Wang, B., Zhang, Y., Cui, Q., Kang, F., Jiang, B., An, M., Ren, Y., et al.: Matrix-game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009 (2025)
33. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
34. Hong, Y., Mei, Y., Ge, C., Xu, Y., Zhou, Y., Bi, S., Hold-Geoffroy, Y., Roberts, M., Fisher, M., Shechtman, E., et al.: Relic: Interactive video world model with long-horizon memory. arXiv preprint arXiv:2512.04040 (2025)
35. Hong, Y., Liu, B., Wu, M., Zhai, Y., Chang, K.W., Li, L., Lin, K., Lin, C.C., Wang, J., Yang, Z., et al.: Slowfast-vgen: Slow-fast learning for action-driven long video generation. arXiv preprint arXiv:2410.23277 (2024)
36. Hou, C., Wei, G., Zeng, Y., Chen, Z.: Training-free camera control for video generation. arXiv preprint arXiv:2406.10126 (2024)
37. Hou, Y., Zheng, L., Torr, P.: Learning camera movement control from real-world drone videos. arXiv preprint arXiv:2412.09620 (2024)
38. Hu, T., Peng, H., Liu, X., Ma, Y.: Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. arXiv preprint arXiv:2506.05554 (2025)
39. Huang, T., Zheng, W., Wang, T., Liu, Y., Wang, Z., Wu, J., Jiang, J., Li, H., Lau, R., Zuo, W., et al.: Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. ACM Transactions on Graphics (TOG) **44**(6), 1–15 (2025)
40. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
41. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009 (2025)
42. Huang, Z., Jeong, H., Chen, X., Gryaditskaya, Y., Wang, T.Y., Lasenby, J., Huang, C.H.: Spacetimepilot: Generative rendering of dynamic scenes across space and time. arXiv preprint arXiv:2512.25075 (2025)
43. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21807–21818 (2024)
44. Jin, W., Dai, Q., Luo, C., Baek, S.H., Cho, S.: Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. arXiv preprint arXiv:2502.08244 (2025)
45. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
46. Lee, J.H., Lin, B.J., Sun, W.F., Lee, C.Y.: Enhancing memory and imagination consistency in diffusion-based world models via linear-time sequence modeling. arXiv e-prints pp. arXiv–2502 (2025)

47. Lee, J., Jung, J., Han, J., Narihira, T., Fukuda, K., Seo, J., Hong, S., Mitsufuji, Y., Kim, S.: 3d scene prompting for scene-consistent camera-controllable video generation. arXiv preprint arXiv:2510.14945 (2025)
48. Li, G., Zheng, S., Xu, S., Chen, J., Li, B., Hu, X., Zhao, L., Jiang, P.T.: Magicworld: Interactive geometry-driven video world exploration. arXiv preprint arXiv:2511.18886 (2025)
49. Li, L., Zhang, Z., Li, Y., Xu, J., Hu, W., Li, X., Cheng, W., Gu, J., Xue, T., Shan, Y.: Nvcomposer: Boosting generative novel view synthesis with multiple sparse and unposed images. arXiv preprint arXiv:2412.03517 (2024)
50. Li, R., Torr, P., Vedaldi, A., Jakab, T.: Vmem: Consistent interactive video scene generation with surfel-indexed view memory. arXiv preprint arXiv:2506.18903 (2025)
51. Li, T., Zheng, G., Jiang, R., Wu, T., Lu, Y., Lin, Y., Li, X., et al.: Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. arXiv preprint arXiv:2502.10059 (2025)
52. Li, Y., Wang, S., Guo, M., Huang, J., Ding, T., Hu, M., Wang, K., Shen, S., Tan, G.: Recamdriving: Lidar-free camera-controlled novel trajectory video generation. arXiv preprint arXiv:2512.03621 (2025)
53. Li, Z., Yu, H.X., Liu, W., Yang, Y., Herrmann, C., Wetzstein, G., Wu, J.: Wonderplay: Dynamic 3d scene generation from a single image and actions. In: Proceedings of the IEEE/CVF international conference on computer vision (2025)
54. Liang, H., Cao, J., Goel, V., Qian, G., Korolev, S., Terzopoulos, D., Plataniotis, K.N., Tulyakov, S., Ren, J.: Wonderland: Navigating 3d scenes from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 798–810 (2025)
55. Liao, K., Wu, S., Wu, Z., Jin, L., Wang, C., Wang, Y., Wang, F., Li, W., Loy, C.C.: Thinking with camera: A unified multimodal model for camera-centric understanding and generation. arXiv preprint arXiv:2510.08673 (2025)
56. Ling, L., Sheng, Y., Tu, Z., Zhao, W., Xin, C., Wan, K., Yu, L., Guo, Q., Yu, Z., Lu, Y., et al.: Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22160–22169 (2024)
57. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: The Eleventh International Conference on Learning Representations (2023)
58. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: ReconX: reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024)
59. Liu, K., Hu, W., Xu, J., Shan, Y., Lu, S.: Rolling forcing: Autoregressive long video diffusion in real time. arXiv preprint arXiv:2509.25161 (2025)
60. Liu, P., Guo, Z., Warke, M., Chintala, S., Paxton, C., Shafiullah, N.M.M., Pinto, L.: Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 13346–13355. IEEE (2025)
61. Lu, D., Liang, A., Huang, T., Fu, X., Zhao, Y., Ma, B., Pan, L., Yin, W., Kong, L., Ooi, W.T., et al.: See4d: Pose-free 4d generation via auto-regressive video inpainting. arXiv preprint arXiv:2510.26796 (2025)
62. Luo, Y., Shi, X., Bai, J., Xia, M., Xue, T., Wang, X., Wan, P., Zhang, D., Gai, K.: Camclonemaster: Enabling reference-based camera control for video generation. In: Proceedings of the SIGGRAPH Asia 2025 Conference Papers. pp. 1–10 (2025)

63. Ma, Y., Zheng, X., Xu, J., Xu, X., Ling, F., Zheng, X., Kuang, H., Li, H., Wang, X., Xiao, X., et al.: Flow caching for autoregressive video generation. arXiv preprint arXiv:2602.10825 (2026)
64. Mao, X., Li, Z., Li, C., Xu, X., Ying, K., He, T., Pang, J., Qiao, Y., Zhang, K.: Yume-1.5: A text-controlled interactive world generation model. arXiv preprint arXiv:2512.22096 (2025)
65. Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulò, S.R., Nießner, M., Kontschieder, P.: Multidiff: Consistent novel view synthesis from a single image. In: Proc. CVPR (2024)
66. Pan, P., Lin, C., Zhao, J., Li, C., Lin, Y., Li, H., Yan, H., Wen, K., Lin, Y., Yuan, Y., et al.: Diff4splat: Controllable 4d scene generation with latent dynamic reconstruction models. arXiv preprint arXiv:2511.00503 (2025)
67. Park, J., Kwon, T., Ye, J.C.: Zero4d: Training-free 4d video generation from single video using off-the-shelf video diffusion. arXiv preprint arXiv:2503.22622 (2025)
68. Parker-Holder, J., Ball, P., Bruce, J., Dasagi, V., Holsheimer, K., Kaplanis, C., Moufarek, A., Scully, G., Shar, J., Shi, J., Spencer, S., Yung, J., Dennis, M., Kenjeyev, S., Long, S., Mnih, V., Chan, H., Gazeau, M., Li, B., Pardo, F., Wang, L., Zhang, L., Besse, F., Harley, T., Mitenkova, A., Wang, J., Clune, J., Hassabis, D., Hadsell, R., Bolton, A., Singh, S., Rocktäschel, T.: Genie 2: A large-scale foundation world model. Google DeepMind (2024), https://deepmind.google/discover/blog/genie-2-a-large-scale-foundation-world-model/
69. Po, R., Nitzan, Y., Zhang, R., Chen, B., Dao, T., Shechtman, E., Wetzstein, G., Huang, X.: Long-context state-space video world models. arXiv preprint arXiv:2505.20171 (2025)
70. Popov, S., Raj, A., Krainin, M., Li, Y., Freeman, W.T., Rubinstein, M.: Camctrl3d: Single-image scene exploration with precise 3d camera control. arXiv preprint arXiv:2501.06006 (2025)
71. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025)
72. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
73. Samuel, D., Tzachor, I., Levy, M., Green, M., Chechik, G., Ben-Ari, R.: Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. arXiv preprint arXiv:2602.01801 (2026)
74. Savov, N., Kazemi, N., Zhang, D., Paudel, D.P., Wang, X., Van Gool, L.: Statespacediffuser: Bringing long context to diffusion world models. arXiv preprint arXiv:2505.22246 (2025)
75. Seo, J., Fukuda, K., Shibuya, T., Narihira, T., Murata, N., Hu, S., Lai, C.H., Kim, S., Mitsufuji, Y.: Genwarp: Single image to novel views with semantic-preserving generative warping. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024)
76. Song, C., Yang, Y., Zhao, T., Li, R., Zhang, C.: Worldforge: Unlocking emergent 3d/4d generation in video diffusion model via training-free guidance. arXiv preprint arXiv:2509.15130 (2025)
77. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. arXiv preprint arXiv:2502.06764 (2025)

78. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928 (2024)
79. Sun, W., Zhang, H., Wang, H., Wu, J., Wang, Z., Wang, Z., Wang, Y., Zhang, J., Wang, T., Guo, C.: Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. arXiv preprint arXiv:2512.14614 (2025)
80. Team, R., Gao, Z., Wang, Q., Zeng, Y., Zhu, J., Cheng, K.L., Li, Y., Wang, H., Xu, Y., Ma, S., et al.: Advancing open-source world models. arXiv preprint arXiv:2601.20540 (2026)
81. Van Hoorick, B., Chen, D., Iwase, S., Tokmakov, P., Irshad, M.Z., Vasiljevic, I., Gupta, S., Cheng, F., Zakharov, S., Guizilini, V.C.: Anyview: Synthesizing any novel view in dynamic scenes. arXiv preprint arXiv:2601.16982 (2026)
82. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: European Conference on Computer Vision. pp. 313–331. Springer (2024)
83. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
84. Wang, C., Zhuang, P., Ngo, T.D., Menapace, W., Siarohin, A., Vasilkovsky, M., Skorokhodov, I., Tulyakov, S., Wonka, P., Lee, H.Y.: 4real-video: Learning generalizable photo-realistic 4d video diffusion. arXiv preprint arXiv:2412.04462 (2024)
85. Wang, Z., Wang, T., Zhang, H., Zuo, X., Wu, J., Wang, H., Sun, W., Wang, Z., Cao, C., Zhao, H., et al.: Worldcompass: Reinforcement learning for long-horizon world models. arXiv preprint arXiv:2602.09022 (2026)
86. Wang, Z., Xia, X., Bie, Z., Liu, J., Yu, D., Bian, J.W., Wang, C.: Taming camera-controlled video generation with verifiable geometry reward. arXiv preprint arXiv:2512.02870 (2025)
87. Wang, Z., Yuan, Z., Wang, X., Li, Y., Chen, T., Xia, M., Luo, P., Shan, Y.: Motionctrl: A unified and flexible motion controller for video generation. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–11 (2024)
88. Wang, Z., Cho, J., Li, J., Lin, H., Yoon, J., Zhang, Y., Bansal, M.: Epic: Efficient video camera control learning with precise anchor-video guidance. arXiv preprint arXiv:2505.21876 (2025)
89. Watson, D., Saxena, S., Li, L., Tagliasacchi, A., Fleet, D.J.: Controlling space and time with diffusion models. In: The Thirteenth International Conference on Learning Representations (2024)
90. Wu, H., Wu, D., He, T., Guo, J., Ye, Y., Duan, Y., Bian, J.: Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982 (2025)
91. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. Proc. CVPR (2025)

92. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory (2025), https://arxiv.org/abs/2506.05284
93. Wu, X., Zhang, G., Xu, Z., Zhou, Y., Lu, Q., He, X.: Pack and force your memory: Long-form and consistent video generation. arXiv preprint arXiv:2510.01784 (2025)
94. Xiang, J., Gu, Y., Liu, Z., Feng, Z., Gao, Q., Hu, Y., Huang, B., Liu, G., Yang, Y., Zhou, K., et al.: Pan: A world model for general, interactable, and long-horizon world simulation. arXiv preprint arXiv:2511.09057 (2025)
95. Xiao, Z., Lan, Y., Zhou, Y., Ouyang, W., Yang, S., Zeng, Y., Pan, X.: Worldmem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369 (2025)
96. Xiao, Z., Ouyang, W., Zhou, Y., Yang, S., Yang, L., Si, J., Pan, X.: Trajectory attention for fine-grained video motion control. arXiv preprint arXiv:2411.19324 (2024)
97. Xiao, Z., Zhao, Y., Li, L., Lan, Y., Yu, N., Garg, R., Cooper, R., Taghavi, M.H., Pan, X.: Video4spatial: Towards visuospatial intelligence with context-guided video generation. arXiv preprint arXiv:2512.03040 (2025)
98. Xie, M., Khan, N., Wang, T., Dhingra, N., Nam, S., Yang, H., Hui, Z., Metzler, C., Vedaldi, A., Pirsiavash, H., et al.: Lavr: Scene latent conditioned generative video trajectory re-rendering using large 4d reconstruction models. arXiv preprint arXiv:2601.14674 (2026)
99. Xu, D., Nie, W., Liu, C., Liu, S., Kautz, J., Wang, Z., Vahdat, A.: Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv preprint arXiv:2406.02509 (2024)
100. Yang, S., Huang, W., Chu, R., Xiao, Y., Zhao, Y., Wang, X., Li, M., Xie, E., Chen, Y., Lu, Y., et al.: Longlive: Real-time interactive long video generation. arXiv preprint arXiv:2509.22622 (2025)
101. Yang, X., Xu, J., Luan, K., Zhan, X., Qiu, H., Shi, S., Li, H., Yang, S., Zhang, L., Yu, C., et al.: Omnicam: Unified multimodal video generation via camera control. arXiv preprint arXiv:2504.02312 (2025)
102. Yang, Y., Fan, L., Shi, Z., Peng, J., Wang, F., Zhang, Z.: Neoverse: Enhancing 4d world model with in-the-wild monocular videos. arXiv preprint arXiv:2601.00393 (2026)
103. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
104. Ye, D., Zhou, F., Lv, J., Ma, J., Zhang, J., Lv, J., Li, J., Deng, M., Yang, M., Fu, Q., et al.: Yan: Foundational interactive video generation. arXiv preprint arXiv:2508.08601 (2025)
105. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. arXiv preprint arXiv:2405.15364 (2024)
106. Yu, H., Wang, C., Zhuang, P., Menapace, W., Siarohin, A., Cao, J., Jeni, L., Tulyakov, S., Lee, H.Y.: 4real: Towards photorealistic 4d scene generation via video diffusion models. Advances in Neural Information Processing Systems **37**, 45256–45280 (2024)
107. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: Wonderworld: Interactive 3d scene generation from a single image. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5916–5926 (2025)

108. Yu, J., Bai, J., Qin, Y., Liu, Q., Wang, X., Wan, P., Zhang, D., Liu, X.: 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)
109. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint arXiv:2503.05638 (2025)
110. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024)
111. Yu, W., Yin, S., Easterbrook, S., Garg, A.: Egosim: Egocentric exploration in virtual worlds with multi-modal conditioning. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=zAyS5aRKV8
112. Zhang, C., Li, B., Wei, M., Cao, Y.P., Gambardella, C.C., Phung, D., Cai, J.: Unified camera positional encoding for controlled video generation. arXiv preprint arXiv:2512.07237 (2025)
113. Zhang, D.J., Paiss, R., Zada, S., Karnad, N., Jacobs, D.E., Pritch, Y., Mosseri, I., Shou, M.Z., Wadhwa, N., Ruiz, N.: Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. arXiv preprint arXiv:2411.05003 (2024)
114. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025)
115. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3d modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21685–21695 (2025)
116. Zhang, T., Bi, S., Hong, Y., Zhang, K., Luan, F., Yang, S., Sunkavalli, K., Freeman, W.T., Tan, H.: Test-time training done right. arXiv preprint arXiv:2505.23884 (2025)
117. Zhang, Z., Chen, D., Liao, J.: I2v3d: Controllable image-to-video generation with 3d guidance. arXiv preprint arXiv:2503.09733 (2025)
118. Zhao, J., Wei, F., Liu, Z., Zhang, H., Xu, C., Lu, Y.: Spatia: Video generation with updatable spatial memory. arXiv preprint arXiv:2512.15716 (2025)
119. Zheng, G., Li, T., Jiang, R., Lu, Y., Wu, T., Li, X.: Cami2v: Camera-controlled image-to-video diffusion model. arXiv preprint arXiv:2410.15957 (2024)
120. Zheng, S., Peng, Z., Zhou, Y., Zhu, Y., Xu, H., Huang, X., Fu, Y.: Vidcraft3: Camera, object, and lighting control for image-to-video generation. arXiv preprint arXiv:2502.07531 (2025)
121. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv e-prints pp. arXiv–2503 (2025)
122. Zhou, S., Du, Y., Yang, Y., Han, L., Chen, P., Yeung, D.Y., Gan, C.: Learning 3d persistent embodied world models. arXiv preprint arXiv:2505.05495 (2025)
123. Zhou, T., Tucker, R., Flynn, J., Fyffe, G., Snavely, N.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)
124. Zhou, Z., An, J., Luo, J.: Latent-reframe: Enabling camera control for video diffusion model without training. arXiv preprint arXiv:2412.06029 (2024)