

DCARL: A Divide-and-Conquer Framework for Autoregressive Long-Trajectory Video Generation

Junyi Ouyang^{1,2}, Wenbin Teng^{1,2}, Gonglin Chen^{1,2}, Yajie Zhao^{1,2}, and Haiwei Chen^{1,2}

¹ Institute for Creative Technologies, USA

² University of Southern California, USA

{junyiouy, wenbinte, gonglinc}@usc.edu, zhao@ict.usc.edu,
chw9308@hotmail.com

Abstract. Long-trajectory video generation is a crucial yet challenging task for world modeling, primarily due to the limited scalability of existing video diffusion models (VDMs). Autoregressive models, while offering infinite rollout, suffer from visual drift and poor controllability. To address these issues, we propose *DCARL*, a novel divide-and-conquer, autoregressive framework that effectively combines the structural stability of the divide-and-conquer scheme with the high-fidelity generation of VDMs. Our approach first employs a dedicated Keyframe Generator trained without temporal compression to establish long-range, globally consistent structural anchors. Subsequently, an Interpolation Generator synthesizes the dense frames in an autoregressive manner with overlapping segments, utilizing the keyframes for global context and a single clean preceding frame for local coherence. Trained on a large-scale internet long trajectory video dataset, our method achieves superior performance in both visual quality (lower FID and FVD) and camera adherence (lower ATE and ARE) compared to state-of-the-art autoregressive and divide-and-conquer baselines, demonstrating stable and high-fidelity generation for long trajectory videos up to 32 seconds in length.

Keywords: Long-trajectory video generation, Video diffusion models, Controllable video synthesis

1 Introduction

Long trajectory video generation has emerged as a pivotal research direction in recent years [6–8, 18, 22, 28, 35]. Developing generative models that can simulate large-scale, realistic scenes is of paramount importance for terrain reconstruction, autonomous driving systems and world models. To achieve this, existing methods [1, 9, 10, 34] typically leverage video diffusion models (VDMs) to facilitate high-fidelity generation. However, while these models perform well on room-sized scenes, they scale poorly to the synthesis of extended sequences (e.g., 30-second videos). This is primarily because predicting the entire sequence collectively incurs a prohibitive computational cost, limiting their effectiveness for long-duration and large-scene tasks. Autoregressive (AR) diffusion models [3, 12, 15]

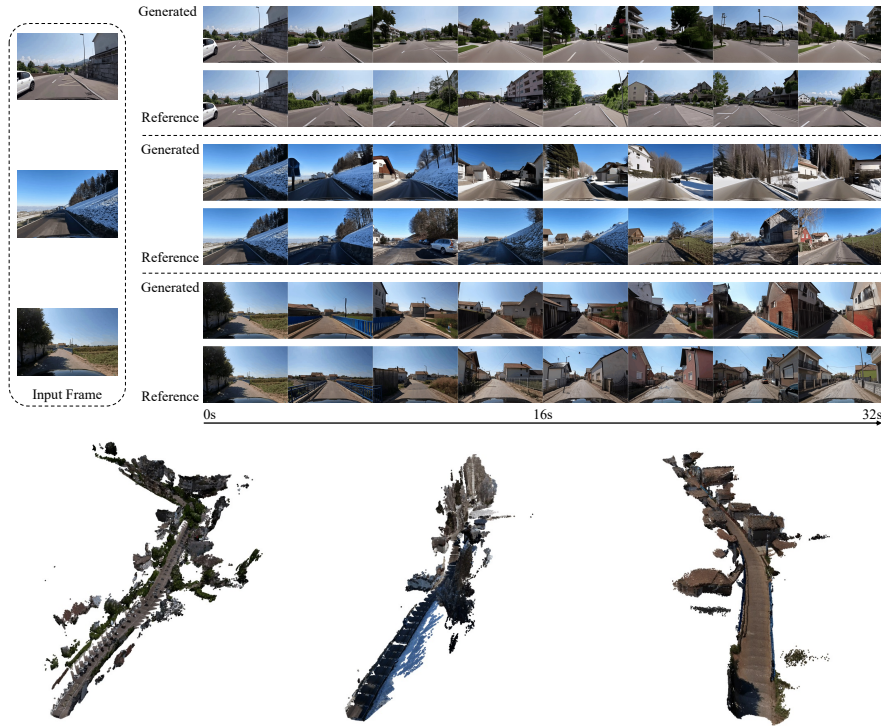


Fig. 1: Given a single *Input Frame* (left), a camera trajectory and a text description (omitted for brevity), *DCARL* generates **high-fidelity** and **temporally consistent** future scene sequences spanning 32s. The figure showcases synthesis results across three diverse environments: urban (top), snowy suburban (middle), and rural (bottom). The bottom panel shows the corresponding 3D point cloud reconstructions by π_3 [24] of the generated sequences. The *Generated* sequences exhibit strong visual realism and long-term temporal coherence that closely match the quality of *Reference* sequences, highlighting our method’s efficacy for long-trajectory generation.

introduce causal attention mechanisms to VDMs, enabling long-sequence roll-out through iterative next-token prediction. However, these models often suffer from severe drift due to exposure bias—a phenomenon arising from the discrepancy between training on ground-truth contexts and inferring from self-generated frames [22, 36]. These compounding prediction errors not only degrade visual quality but also severely compromise the effectiveness of external control signals like camera poses. As the synthesized content deviates from the valid manifold, the model increasingly loses its ability to interpret and adhere to the geometric constraints provided by the control trajectory. This decoupling creates a fundamental misalignment between the intended camera instructions and the actual synthesized motion, ultimately causing purely AR methods to fail in maintaining trajectory consistency over long trajectories.

Existing methods seek to alleviate exposure bias by mitigating the training–inference mismatch. Diffusion Forcing [3, 19] injects random Gaussian noise into conditional inputs, while Self Forcing [12] trains causal models on self-generated rollouts to narrow the distribution gap. Although these strategies mitigate local drifting, they fail to explicitly bound error propagation, leaving the challenge of stable frame quality and precise camera control on long trajectories largely unaddressed. Furthermore, techniques [16, 32] that incorporate attention sinks [26] can further impede controllability; fixed sink frames often introduce sink collapse [5] and severely hinder the model’s responsiveness to dynamic pose instructions, ultimately degrading camera motion fidelity and scene transitions across extended trajectories.

Therefore, in this work, we propose *DCARL*, a novel framework that utilizes VDM in an autoregressive paradigm. Our method follows a divide-and-conquer scheme, employing dual generators specifically tailored for the synthesis of global keyframes and local dense frames. Our key insight is that purely autoregressive generation inherently suffers from unbounded error accumulation. To overcome this, our framework first employs a full-sequence generator to jointly synthesize sparse keyframes across the entire video timeline. By modeling these global keyframes jointly rather than sequentially, we effectively bypass autoregressive error propagation, establishing structurally accurate and semantic-consistent anchors for the entire long sequence.

Building upon these global anchors, our dense frame generator synthesizes the intermediate video segments via a constrained interpolation process. Specifically, the local generation is conditioned on both the surrounding keyframes, which serve as fixed constraints, and the preceding dense segment. We identify two failure modes in keyframe-interpolation-based video generation. First, the model can become overly anchored to the provided keyframes, causing keyframe duplication in nearby frames rather than smooth, physically plausible motion, yielding unnatural dynamics. The second is boundary inconsistency, where the video exhibits abrupt transitions at segment boundaries. To address these issues, we propose *Motion-Inductive Noisy Conditioning* and a *Seamless Boundary Consistency* module, enabling more dynamic and natural long-video generation. As validated by our empirical results, our modeling of global keyframe anchors and local transitions achieves highly stable generation of high-fidelity, long trajectory videos with precise camera control. Notably, our approach successfully mitigates the severe visual drifts lie in existing autoregressive approaches, enabling controllable long trajectory video generation. Fig. 1 illustrates representative long-trajectory generation results produced by our framework.

2 Related Works

Autoregressive Video Generation. Autoregressive video generation models synthesize a long sequence in a causal manner by predicting the next frame (or block) conditioned on past generated content. Therefore, they naturally support streaming output and long-horizon rollout, but they also amplify exposure bias,

error accumulation, and long-context constraints. Recent works replace purely teacher-forced learning with noisy or self-conditioned rollouts to reduce the train-test mismatch, and modify positional embeddings to stabilize long-context generation [3, 4, 12, 16, 19, 30–32, 38]. Despite strong small-scene performance, these methods remain challenged in long-horizon scene generation, where rapid scene changes exacerbate the difficulties of pure autoregressive generation and controllability.

Keyframe-Interpolation Video Generation. An alternative is coarse-to-fine keyframe interpolation: sparse keyframes provide global structural anchors, and an interpolation model synthesizes intervening motion. Diffusion-based models generate intermediate frames from endpoint keyframes [14, 23]. For longer sequences, NUWA-XL [33] and SEVA [37] use hierarchical keyframe-based generation to extend video horizons. However, they rely on a single generator to jointly model global keyframes and local dense synthesis, coupling the distinct distributions of long-range structural planning and short-range interpolation. This coupling makes it difficult to preserve global camera-conditioned structure while maintaining local temporal continuity over long trajectories.

Long Horizon Scene Generation. Long-horizon scene generation seeks controllable, temporally consistent rollouts over hundreds or thousands of frames, challenged by compounding drift, long-context scaling, and multi-view consistency. Recent diffusion world models aim to make long rollouts more stable and controllable: VISTA [8] and GAIA [18] emphasize richer conditional interfaces and design choices that better preserve structure across iterative sampling, while Epona [35] explicitly frames long-horizon synthesis as autoregressive diffusion and trains the model to tolerate its own rollout errors. However, these methods still rely on autoregressive generation for long videos and remain susceptible to quality degradation during rollout.

3 Preliminary

Autoregressive Generation. Long-horizon video generation is typically formulated as an iterative factorization of the joint distribution conditioned on the control signals $\mathcal{C}$:

$$p(\mathcal{V}|\mathcal{C}) = \prod_{i=1}^S p(S_i|S_{<i}, \mathcal{C}), \quad (1)$$

where the sequence $\mathcal{V}$ is partitioned into S contiguous blocks $\{S_1, \dots, S_S\}$ of size B , which controls the generation granularity. A smaller B lowers per-step cost but requires more autoregressive updates, whereas a larger B increases per-step cost while expanding the bidirectional receptive field. Thus, B trades off per-step efficiency, sequential depth, and local consistency.

Divide-and-Conquer Paradigm. An alternative strategy for long-video synthesis involves a hierarchical decomposition of the generation task into global structural planning and local detail refinement. This paradigm is typically formulated by introducing a sparse structural proxy $\mathcal{K}$, such as keyframes or semantic

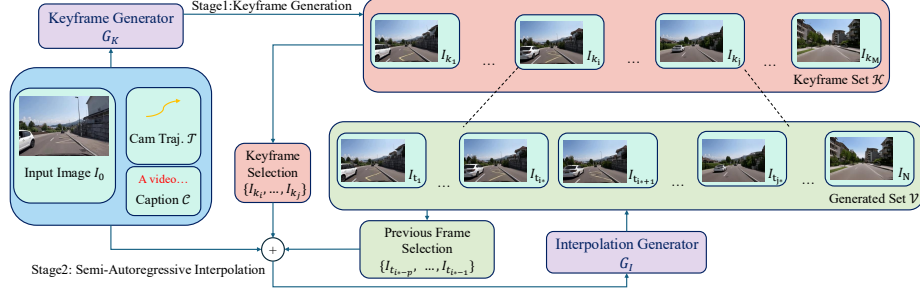


Fig. 2: Our Two-Stage Autoregressive Pipeline for Long-Term Video Synthesis. Our model synthesizes a long video sequence $\mathcal{V} = \{I_1, \dots, I_N\}$ conditioned on multimodal inputs: the initial image I_0 , camera trajectory $\mathcal{T}$, and caption $\mathcal{C}$. **The Keyframe Generator** G_K produces a global Keyframe Set $\mathcal{K}$, establishing the overall structural evolution of the video. **The Interpolation Generator** G_I generates the final frames in an autoregressive manner.

layouts, which abstracts the temporal evolution of the sequence:

$$p(\mathcal{V}|\mathcal{C}) = p(\mathcal{K}|\mathcal{C})p(\mathcal{V}|\mathcal{K}, \mathcal{C}). \quad (2)$$

Establishing the global trajectory via $\mathcal{K}$ transforms dense synthesis into a constrained interpolation task. This hierarchy bounds prediction errors within structural anchors, preventing the unbounded drift typical of autoregressive rollout. While decoupling long-range structure from local synthesis enables scalable generation, the paradigm remains sensitive to the noise inherent in the structural proxy $\mathcal{K}$ and requires high-fidelity conditional mapping in $p(\mathcal{V}|\mathcal{K}, \mathcal{C})$ to maintain overall consistency.

4 Methods

4.1 Overview

We define the task of long-trajectory scene generation as generating a dense video sequence $\mathcal{V} = \{I_1, I_2, \dots, I_N\}$ of length N , conditioned on an initial frame I_0 and a sequence of control signals (e.g., camera poses) $\mathcal{T} = \{c_1, c_2, \dots, c_N\}$. The goal is to model the conditional distribution $p(\mathcal{V}|I_0, \mathcal{T})$.

As illustrated in Fig. 2, our framework follows a two-stage divide-and-conquer pipeline. In the first stage, the keyframe generator G_K synthesizes a sparse keyframe set $\mathcal{K} = \{I_{k_1}, \dots, I_{k_M}\}$ conditioned on the initial image I_0 , camera trajectory $\mathcal{T}$, and caption $\mathcal{C}$. These keyframes establish the overall structural evolution of the video. In the second stage, the interpolation generator G_I synthesizes the dense sequence $\mathcal{V}$ in a segment-wise manner.

Following the autoregressive paradigm, G_I synthesizes each segment S_i by generating its constituent frames simultaneously. The generation is conditioned

on a selected subset of keyframes $\{I_{k_i}, \dots, I_{k_j}\}$ and a history of p previously generated frames $\{I_{t_i^*-p}, \dots, I_{t_i^*-1}\}$ with t_i^* being the first frame index of the current segment. This design decouples global structural planning from local detail interpolation, effectively reducing cumulative drift while maintaining precise adherence to the given trajectory.

4.2 Keyframe Generation

The keyframe generator G_K generates the keyframe set $\mathcal{K}$ as defined in Sec. 4.1. Specifically, we implement this mapping through a DiT-based [17] flow matching process Φ_K :

$$\mathcal{K} = \mathcal{D}(\Phi_K(\epsilon \mid \mathcal{E}(I_0), \mathcal{P}, \mathcal{C})), \quad (3)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the initial Gaussian noise. Here, $\mathcal{E}$ and $\mathcal{D}$ denote the VAE encoder and decoder, and $\mathcal{P} = \{c_{k_1}, \dots, c_{k_M}\}$ represents the sampled poses from the global trajectory $\mathcal{T}$. This process can be autoregressive, where later keyframes are conditioned on earlier generated keyframes. Our insight is generating keyframes would achieve bounded or mitigated error accumulation as illustrated by the following proposition:

Proposition 1 (Temporal Degradation via Perturbation Analysis). *Let $e_t = I_t - I_t^*$ denote the cumulative generation error at frame t , where I_t and I_t^* are the generated and ground-truth frames. We model autoregressive (AR) generation as $I_{t+1} = \mathcal{F}(I_t, \epsilon_t)$, where $\mathcal{F}$ is the generator mapping and ϵ_t is the sampling noise. Assuming $\mathcal{F}$ is L -Lipschitz continuous with respect to the state ($L = \sup \|\nabla_I \mathcal{F}\| \geq 1$), a first-order perturbation analysis yields the cumulative error bound $\|e_N^{base}\| \leq \sum_{j=1}^N L^{N-j} \|\eta_j\|$, where η_j represents the local injection error at step j . If $\|\eta_j\|$ is bounded by a constant η , this cumulative error diverges at least linearly $\mathcal{O}(N)$ (and exponentially if $L > 1$). Our divide-and-conquer strategy mitigates this via keyframe anchoring; the interpolation error at any frame $\|e_t^{ours}\|$ is bounded by the convex combination of its anchoring keyframe errors plus a local constant.*

Please refer to the supplementary material for complete proof. To ensure structural integrity and control precision, we incorporate the following designs: **Spatial-Structural Preservation.** Standard temporal VAEs typically employ temporal compression, which is inherently lossy and degrades visual quality for sparse sequences with significant spatial displacement. To circumvent this, we treat the M keyframes as a batch of independent images by reshaping the frame dimension into the batch dimension during encoding and decoding. This ensures that each generated keyframe maintains full spatial fidelity, providing a reliable foundation for subsequent interpolation.

Keyframe Sampling Strategy. We construct the structural skeleton by sampling keyframes from the sequence using a temporal stride Δk . To ensure robustness across diverse motion scales, Δk is randomly drawn from a discrete set of candidate strides during training. During testing, a constant stride Δk_{test} is maintained. The selection of Δk_{test} serves as a critical hyperparameter that

modulates the trade-off between the computational efficiency of the keyframe generation process and the precision of the structural guidance provided to the interpolation model.

4.3 Interpolation Generation

The interpolation generator G_I synthesizes the dense video sequence $\mathcal{V}$ in a segment-wise manner. Specifically, for each segment S_i , we construct an input latent sequence $\mathbf{z}_{in}$ by concatenating the noise with temporal context. The DiT-based flow matching process Φ_I is formulated as:

$$\begin{aligned} S_i &= G_I(\mathcal{H}_i, \mathcal{K}_i, \mathcal{T}_i, \mathcal{C}) \\ &= \mathcal{D}(\Phi_I(\mathbf{z}_{in}, \mathcal{T}_i, \mathcal{C})), \end{aligned} \quad (4)$$

where $\mathbf{z}_{in} = [\epsilon, \mathbf{z}_{\mathcal{H}_i}, \mathbf{z}_{\mathcal{K}_i}]$ represents the sequence formed by concatenating the initial Gaussian noise ϵ , the encoded history $\mathbf{z}_{\mathcal{H}_i} = \mathcal{E}(\mathcal{H}_i)$, and the selected keyframes $\mathbf{z}_{\mathcal{K}_i} = \mathcal{E}(\mathcal{K}_i)$ along the temporal dimension.

Keyframe Selection Policy. For each segment $S_i = \{I_{t_i^*}, \dots, I_{t_j^*}\}$ and the corresponding trajectory $\mathcal{T}_i = \{c_{t_i^*}, \dots, c_{t_j^*}\}$ in the interpolation stage, we select a local subset of keyframes $\mathcal{K}_i \subset \mathcal{K}$ to provide temporal context. This subset encompasses all keyframes within the segment’s temporal span, along with the nearest keyframes that immediately precede and follow the segment:

$$\mathcal{K}_i = \{I_k \in \mathcal{K} \mid k_{pre} \leq k \leq k_{next}\}, \quad (5)$$

where $k_{pre} = \max\{k \in \mathcal{I}_K \mid k < t_i^*\}$ and $k_{next} = \min\{k \in \mathcal{I}_K \mid k > t_j^*\}$. This policy ensures that each segment is synthesized with both look-back and look-ahead structural constraints, preventing trajectory drift at the segment boundaries.

Autoregressive Rollout. To synthesize long-trajectory videos, G_I operates via a sliding window. During each sampling step of the flow matching process, the noisy latent $\mathbf{z}$ (which originates from ϵ) is continuously updated while being anchored by the fixed historical tokens $\mathbf{z}_{\mathcal{H}_i}$ and global keyframe tokens $\mathbf{z}_{\mathcal{K}_i}$ within $\mathbf{z}_{in}$. This hybrid conditioning allows the model to leverage both local continuity and global anchors in a unified attention space, which effectively mitigates exposure bias and error accumulation compared to purely autoregressive methods, while maintaining superior adherence to the camera trajectory $\mathcal{T}_i$.

Motion-Inductive Noisy Conditioning. A common failure mode in interpolation is keyframe duplication, where the model replicates pixels from keyframes instead of synthesizing motion. This identity mapping shortcut occurs because clean keyframes $\mathbf{z}_{\mathcal{K}_i}$ allow the model to minimize training loss by anchoring to visual guides rather than learning complex temporal transitions, leading to stuttering or unnatural transitions. To mitigate this, we perturb the keyframe latents with a consistent noise level during both training and inference:

$$\tilde{\mathbf{z}}_{\mathcal{K}_i} = \alpha_c \mathbf{z}_{\mathcal{K}_i} + \sigma_c \epsilon, \quad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (6)$$

where the coefficients α_c and σ_c are hyperparameters that control the keyframe guidance and the noise intensity, respectively. This perturbation breaks the pixel-level shortcut, forcing Φ_I to prioritize underlying motion dynamics and trajectory adherence.

Seamless Boundary Consistency. To eliminate visual fractures at segment junctions, we implement a latent substitution strategy consistently during training and inference. Specifically, we define a p -frame overlap between segments and replace the first p latents of segment S_i with historical latents $\mathbf{z}_{\mathcal{H}_i}$ at each denoising step τ . To ensure generalization to these deterministic boundaries, $\mathbf{z}_{\mathcal{H}_i}$ is kept noise-free during Φ_I training. By bridging the distribution gap, this design enforces strict continuity and prevents flickering or motion jumps, ensuring a seamless temporal flow across all segment boundaries.

5 Experiments

5.1 Experiments Setup

Implementation Details. Our model is built upon the Wan2.1-T2V-1.3B [21] architecture. We integrate camera control by adding camera feature maps to visual tokens within each DiT block, following the ReCamMaster [1] design. Both keyframe and interpolation generators are trained for 30,000 steps using the AdamW optimizer (LR: 5×10^{-5} , weight decay: 0.01, β : (0.9, 0.95)) with an effective batch size of 16 across 8 NVIDIA H100 GPUs.

We process a 480-hour subset of the OpenDV-YouTube (ODV-YT) dataset [29], segmented into 1-minute clips at 10 fps. Camera poses are reconstructed using π_3 [24] every 0.5 seconds, with intermediate frames recovered via spherical linear and linear interpolation. Semantic scene descriptions are generated by Qwen2.5-Omni [27] from frames sampled at 20-second intervals.

For keyframe generation, the model is trained with $|K| = 21$ and temporal strides $\Delta k \in \{4, 8, 16\}$, while $\Delta k_{\text{test}} = 8$ is fixed during inference. For the interpolation stage, the number of conditioning keyframes $|K|$ is randomly sampled from $[1, 10]$ during training. In the inference phase, we apply Motion-Inductive Noisy Conditioning with $\alpha_c = 0.7$, $\sigma_c = 0.3$, and set the boundary overlap to $p = 1$. Please refer to the supplementary material for further details.

Baselines. We compare our method with Diffusion Forcing (DiffF) [3], Self Forcing (SelfF) [12], Deep Forcing (DeepF) [32], Vista [8] and Stable Virtual Camera (SEVA) [37]. All models are finetuned for fair comparison.

Datasets. To evaluate the capability of generating videos across large spatial spans and maintain structural integrity in expansive static scenes, we conduct experiments on two large-scale datasets: the ODV-YT validation set and the nuScenes [2] validation set for zero-shot testing. All experiments are conducted on 150 sequences.

Evaluation Metrics. For *video quality*, we use Fréchet Inception Distance (FID) [11] and Fréchet Video Distance (FVD) [20] as primary metrics, reporting both overall and per-segment scores to assess stability. For *camera control*

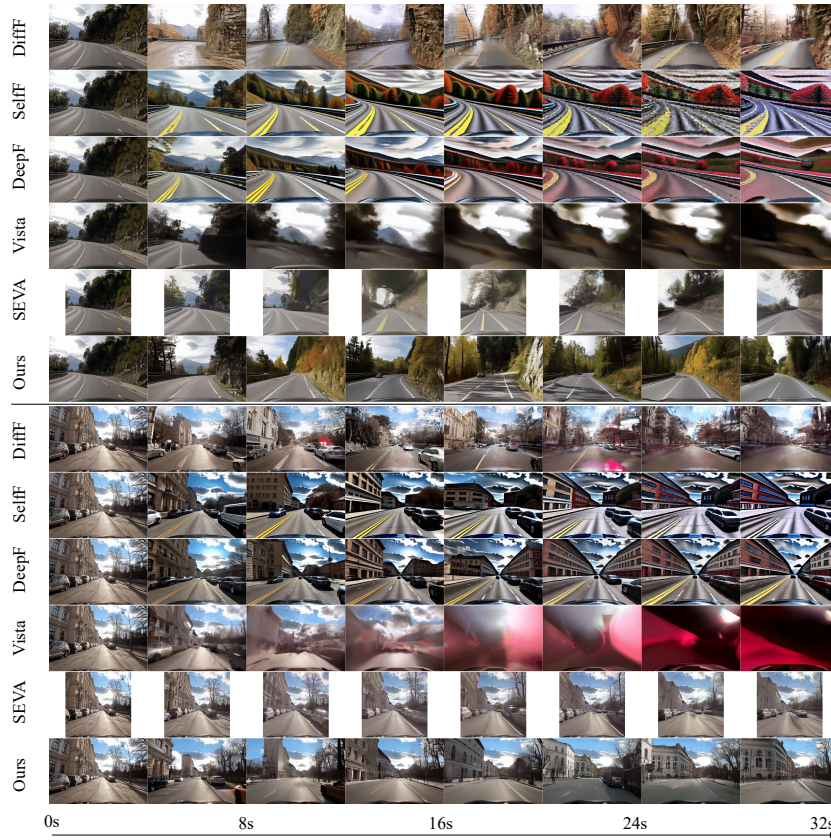


Fig. 3: Qualitative comparison of long-trajectory video generation. We compare our method against several state-of-the-art baselines. Each method is provided with the same initial frame as input. Our approach demonstrates superior long-trajectory consistency and visual fidelity over a 32-second duration.

accuracy, we utilize π_3 [24] to extract 6-DoF camera poses and report Absolute Trajectory Error (ATE) and Absolute Rotation Error (ARE) after alignment. For *segment continuity*, we further use Motion Smoothness (MS) [13], PSNR and SSIM [25] to evaluate the discrepancy between adjacent generated segments.

5.2 Comparison on Long Trajectory Generation.

In Fig. 3, Tab. 1, and Tab. 2, we present qualitative and quantitative comparisons. While autoregressive baseline methods can generate reasonable short-term results, they suffer from severe error accumulation and temporal drifting as the video length increases. Notably, even with the incorporation of specialized strategies designed to mitigate drifting, these methods still exhibit sharp performance degradation on extended sequences. In contrast, our divide-and-conquer

Table 1: Quantitative comparison on 32-second videos of the ODV-YT dataset. Metrics are reported as **FID** / **FVD** for visual quality and temporal continuity, and **ATE** / **ARE** for global path adherence. Our method maintains superior stability and fidelity over extended durations.

Method	Overall FID/FVD↓	0-8 s	8-16 s	16-24 s	24-32 s	ATE ↓	ARE ↓
DiffF	35.0 / 664.1	25.3 / 390.7	39.8 / 684.5	45.2 / 856.8	54.1 / 1067.1	0.469	19.448
SelfF	58.0 / 2113.6	26.2 / 533.5	62.8 / 1911.3	92.8 / 3402.2	119.4 / 4183.8	0.610	14.386
DeepF	42.3 / 1558.5	26.1 / 545.3	49.1 / 1472.3	63.7 / 2342.2	77.2 / 2832.6	0.571	15.144
Vista	66.7 / 1550.0	29.1 / 428.7	66.3 / 1212.6	103.9 / 2452.3	134.1 / 3774.5	0.641	19.332
SEVA	22.2 / 548.0	27.2 / 582.0	30.7 / 643.2	31.8 / 623.8	33.1 / 602.0	0.294	8.527
Ours	19.2 / 203.7	19.6 / 191.4	25.1 / 255.8	27.3 / 276.0	28.6 / 313.8	0.237	7.669

Table 2: Quantitative evaluation of zero-shot generalization on 16-second videos of the nuScenes dataset. Our approach achieves consistent high-fidelity generation and precise camera control in both FVD and ATE.

Method	Overall FID/FVD↓	0-4 s	4-8 s	8-12 s	12-16 s	ATE ↓	ARE ↓
DiffF	37.1 / 566.6	29.3 / 376.3	41.8 / 573.7	51.7 / 738.9	60.9 / 906.2	0.154	13.704
SelfF	53.1 / 711.2	53.1 / 394.8	59.3 / 592.1	69.1 / 989.0	78.7 / 1504.7	0.092	12.464
DeepF	50.3 / 618.4	52.5 / 343.3	59.0 / 577.3	64.7 / 887.2	69.9 / 1174.8	0.104	10.774
Vista	68.7 / 1239.1	27.3 / 445.1	54.2 / 847.0	111.8 / 1860.8	160.5 / 2966.4	0.195	16.607
SEVA	35.9 / 487.9	30.2 / 493.7	41.3 / 529.1	54.7 / 595.2	72.1 / 769.4	0.117	6.289
Ours	19.6 / 225.4	23.3 / 239.9	26.5 / 264.3	31.1 / 323.0	30.8 / 345.2	0.045	5.274

autoregressive design leverages global keyframes as structural anchors to ensure both long-range visual stability and effective camera control. This advantage is further demonstrated by our robust zero-shot generalization to unseen environments. When evaluated on the nuScenes dataset, our approach consistently maintains high generation quality and precise path adherence, significantly surpassing baseline methods without requiring additional fine-tuning. Besides, our method outperforms SEVA in terms of camera controllability and video quality. Quantitatively, these results confirm that our framework remains remarkably stable across the entire horizon, providing superior performance in both in-distribution and out-of-distribution scenarios. For more visualization results, please refer to the supplementary material.

5.3 Ablation Study

Keyframe Spatial-Structural Preservation. We compare our image-level encoding with a temporal compression baseline to evaluate its necessity for sparse keyframe generation. Results are calculated using the generated keyframe sequences for 16-second videos on ODV-YT. As shown in Tab. 3 and Fig. 4, applying temporal compression to keyframes with large spatial spans leads to significant performance drops. Specifically, the spatial displacement between distant keyframes makes the VAE’s 4:1 temporal compression inherently lossy, which severely degrades quality (higher FID). Furthermore, such compression obscures



Fig. 4: Qualitative comparison of keyframe model designs. The top row shows the results of the *temporal compression* baseline, while the bottom row shows our proposed design. Our approach provides significantly clearer visual details and superior structural accuracy compared to the baseline.

fine-grained geometric cues, hindering the camera encoder from learning precise representations, as evidenced by the elevated ATE and ARE.

Table 3: Ablation study on keyframe model design. We compare our proposed image-level encoding against a *temporal compression* baseline, which treats keyframes as a video sequence with temporal downsampling. Our design achieves superior performance in both visual quality and trajectory adherence in keyframe generation.

Model Design	FID↓	ATE↓	ARE↓
Temporal compression	19.6	0.155	7.276
Ours	16.3	0.100	3.999

Keyframe Sampling Strides. We investigate the impact of sampling strategies by varying the generation stride Δk_{gen} and interpolation stride Δk_{int} on 16-second ODV-YT video generation. Specifically, keyframes are initially synthesized at a stride of Δk_{gen} and subsequently filtered to the interpolation stride Δk_{int} to serve as conditioning anchors for the dense synthesis stage. As shown in Tab. 4, Δk_{gen} is the primary determinant of both visual quality and trajectory fidelity. Increasing Δk_{gen} from 4 to 16 consistently enhances the quality of structural anchors, which directly improves the final FID and FVD. In contrast, the interpolation stride Δk_{int} exerts a negligible influence on the overall performance. Regarding trajectory adherence, we find that neither excessively dense nor sparse generation is optimal for ATE and ARE; the former introduces redundant constraints that hinder motion naturalness, while the latter provides insufficient guidance to bound drift. Our balanced configuration of (8, 8) achieves the optimal trade-off between structural stability and generative quality.

Motion-Inductive Noisy Conditioning. Conditioning the interpolation model on noisy keyframes is essential for high-quality synthesis, as discussed in Sec. 4.3. We evaluate this in Tab. 5 and Fig. 5 on 16-second ODV-YT sequences. Quantitatively, we observe that while FID remains relatively stable, the FVD signif-

Table 4: Ablation on keyframe intervals for generation (Δk_{gen}) and interpolation (Δk_{int}). Keyframes are initially generated at stride Δk_{gen} and subsequently filtered to Δk_{int} for dense interpolation.

$(\Delta k_{gen}, \Delta k_{int})$	(4, 4)	(4, 8)	(4, 16)	(8, 8)	(8, 16)	(16, 16)
FID ↓	21.4	21.0	20.9	19.3	19.0	18.8
FVD ↓	201.8	199.9	204.6	196.0	201.5	187.8
ATE ↓	0.148	0.147	0.131	0.096	0.098	0.104
ARE ↓	4.98	4.91	5.08	3.96	4.06	4.74

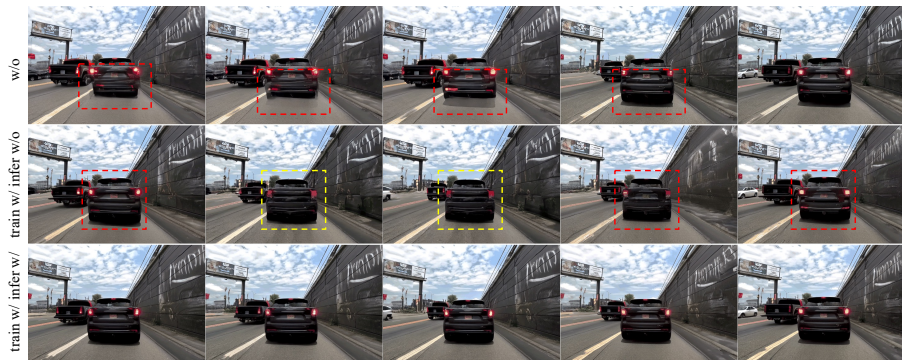


Fig. 5: Ablation on noisy keyframe design. We compare *W/o noisy keyframe*, *Train w/ noisy keyframe*, and our full *Train/infer w/ noisy keyframe* strategy. Applying noise during both stages effectively mitigates memory bias and ensures smoother transitions.

Table 5: Ablation on noisy keyframe design. All methods take the same generated keyframes as input. Applying noisy keyframes during both training and inference achieves the best dynamic video quality (lowest FVD).

Training Design	FID↓	FVD↓
W/o noisy keyframe	19.0	246.5
Train w/ noisy keyframe	19.9	240.9
Train/infer w/ noisy keyframe	19.3	201.7

icantly deteriorates when noise is not consistently applied during both training and inference. Qualitative results in Fig. 5 reveal two distinct artifact patterns: (1) *unnatural transitions*, such as fade-like blending between keyframes, and (2) *motion stagnation*, where the model replicates imperfect conditioning keyframes. These temporal artifacts—which involve mode collapse where the model copies pixels rather than learning motion dynamics—explain why the FVD increases sharply despite the stable FID. This confirms that joint noisy training and inference forces the model to bypass "copy-paste" shortcuts and achieve smoother, more realistic video generation.

Table 6: Ablation on segment continuity design. We compare *Overlap* (simple overlapping), *+Sub* (latent substitution), and our full *+Sub+Train* strategy. Our method effectively bridges segment junctions, achieving the best motion smoothness and structural fidelity.

Design	MS↑	PSNR↑	SSIM↑
Overlap	0.969	18.15	0.534
+ Sub	0.968	17.95	0.524
+ Sub + Train	0.972	18.75	0.562



Fig. 6: Ablation on Frame Continuity Design. The designs test continuity at segment boundaries: **overlap** (simple overlap); **+ sub** (overlap + latent substitution); and **+ sub + train** (our trained continuity mechanism).

Seamless Boundary Consistency. We evaluate the segment consistency mechanism by providing Ground-Truth (GT) frames as boundary conditions and comparing generated frames proximal to the segment junctions against these targets. As shown in Tab. 6 and Fig. 6, simple segment overlap (*Overlap*) fails to guarantee visual coherence; the final frame of a preceding segment often deviates from the required starting pose of the subsequent one, resulting in noticeable motion and temporal fractures. While naively replacing the first frame with the preceding segment’s last latent (*+Sub*) ensures junction continuity, it introduces pronounced artifacts. This stems from a distribution mismatch where the model, expecting noisy inputs, struggles to process a clean and deterministic conditioning frame without explicit consistency training. In contrast, our full strategy (*+Sub+Train*) effectively bridges this gap, achieving superior smoothness and structural fidelity at segment junctions.

Necessity of Keyframe Anchoring. We evaluate the necessity of structural anchors by removing keyframe conditioning during the interpolation stage. Quantitatively, Tab. 7 shows that without keyframe guidance, the model suffers from long-term drifting and poor camera accuracy, with FID/FVD increasing sharply as the sequence extends and ATE/ARE being significantly worse. This confirms that keyframes are essential for bounding accumulated errors over long

Table 7: Ablation study on the effect of keyframes for 32-second videos on the OpenDV-YT dataset. The inclusion of keyframes significantly enhances long-term consistency and trajectory fidelity.

Method	Overall FID/FVD↓	0-8 s	8-16 s	16-24 s	24-32 s	ATE ↓	ARE ↓
w/o Keyframe	25.2 / 376.7	21.1 / 247.1	31.5 / 407.5	38.2 / 546.5	45.2 / 665.3	0.387	12.184
Ours	19.2 / 203.7	19.6 / 191.4	25.1 / 255.8	27.3 / 276.0	28.6 / 313.8	0.237	7.669



Fig. 7: Qualitative ablation of keyframe guidance on 32-second videos. The top row shows the results without keyframe conditioning, and the bottom row shows our full model with keyframe anchors.

durations. Qualitatively, Fig. 7 illustrates that the unguided variant fails to maintain camera control during complex maneuvers, such as a left turn followed by a straight trajectory, resulting in noticeable scene distortion and path deviation. In contrast, our full model leverages keyframe anchors to ensure both visual stability and precise trajectory adherence across the entire 32-second horizon.

6 Conclusion and Future Works

In this paper, we presented *DCARL*, a divide-and-conquer framework that reformulates long-trajectory video generation through an autoregressive modeling paradigm. By decoupling the complex synthesis task into global structural planning via keyframe anchors and local refinement via segment-wise interpolation, our approach effectively bounds the cumulative drift inherent in traditional autoregressive methods. Extensive experiments on the large-scale OpenDV-YouTube and nuScenes datasets demonstrate that *DCARL* consistently achieves state-of-the-art performance in both visual fidelity and trajectory adherence. Notably, our framework enables the stable synthesis of high-quality long trajectories for extended horizons up to 32 seconds, significantly surpassing existing baselines. Despite its effectiveness, several avenues remain for future exploration:

Unified Joint Modeling: While our dual-generator design ensures stability, a promising direction is to investigate unified architectures that can jointly model the distribution of sparse keyframes and dense sequences within a single latent space without compromising generation quality.

Dense Control Priors: While we primarily leverage camera poses for scalable training, integrating richer conditional priors—such as semantic maps, 3D lay-

outs, or LiDAR-based geometries—could further enhance scene complexity and the precision of world-model simulations.

Acknowledgements

Research was sponsored by the Army Research Office and was accomplished under Cooperative Agreement Number W911NF-25-2-0040. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

This material is based upon work supported by the Intelligence Advanced Research Projects Activity under prime Contract No. 140D0423C0034.

References

1. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14834–14844 (2025)
2. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631 (2020)
3. 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)
4. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283* (2025)
5. Cui, J., Wu, J., Li, M., Yang, T., Li, X., Wang, R., Bai, A., Ban, Y., Hsieh, C.J.: Lol: Longer than longer, scaling video generation to hour. *arXiv preprint arXiv:2601.16914* (2026)
6. Gao, R., Chen, K., Xiao, B., Hong, L., Li, Z., Xu, Q.: Magicdrive-v2: High-resolution long video generation for autonomous driving with adaptive control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28135–28144 (2025)
7. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. In: International Conference on Learning Representations. vol. 2024, pp. 22841–22860 (2024)
8. Gao, S., Yang, J., Chen, L., Chitta, K., Qiu, Y., Geiger, A., Zhang, J., Li, H.: Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems* **37**, 91560–91596 (2024)
9. 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)

10. 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13416–13426 (2025)
11. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
12. Huang, X., Li, Z., He, G., Zhou, M., Shechtman, E.: Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems* **38**, 167283–167308 (2026)
13. Huang, Z., Zhang, F., Xu, X., He, Y., Yu, J., Dong, Z., Ma, Q., Chanpaisit, N., Si, C., Jiang, Y., et al.: Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025)
14. Jain, S., Watson, D., Tabellion, E., Poole, B., Kontkanen, J., et al.: Video interpolation with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7341–7351 (2024)
15. Jin, Y., Sun, Z., Li, N., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling. In: International Conference on Learning Representations. vol. 2025, pp. 23378–23402 (2025)
16. 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)
17. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
18. Russell, L., Hu, A., Bertoni, L., Fedoseev, G., Shotton, J., Arani, E., Corrado, G.: Gaia-2: A controllable multi-view generative world model for autonomous driving. *arXiv preprint arXiv:2503.20523* (2025)
19. Song, K., Chen, B., Simchowitz, M., Du, Y., Tedrake, R., Sitzmann, V.: History-guided video diffusion. *arXiv preprint arXiv:2502.06764* (2025)
20. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717* (2018)
21. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314* (2025)
22. Wang, H., Liu, D., Xie, H., Liu, H., Ma, E., Yu, K., Wang, L., Wang, B.: Mila: Multi-view intensive-fidelity long-term video generation world model for autonomous driving. *arXiv preprint arXiv:2503.15875* (2025)
23. Wang, X., Zhou, B., Curless, B., Kemelmacher-Shlizerman, I., Holynski, A., Seitz, S.: Generative inbetweening: Adapting image-to-video models for keyframe interpolation. In: International Conference on Learning Representations. vol. 2025, pp. 30594–30609 (2025)
24. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3 : Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347* (2025)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**(4), 600–612 (2004)

26. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: International Conference on Learning Representations. vol. 2024, pp. 21875–21895 (2024)
27. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., Lin, J.: Qwen2.5-omni technical report (2025)
28. Yan, Y., Xu, Z., Lin, H., Jin, H., Guo, H., Wang, Y., Zhan, K., Lang, X., Bao, H., Zhou, X., et al.: Streetcrafter: Street view synthesis with controllable video diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 822–832 (2025)
29. Yang, J., Gao, S., Qiu, Y., Chen, L., Li, T., Dai, B., Chitta, K., Wu, P., Zeng, J., Luo, P., et al.: Generalized predictive model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14662–14672 (2024)
30. 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)
31. Yesiltepe, H., Meral, T., Akan, A.K., Oktay, K., Yanardag, P.: Infinity-rope: Action-controllable infinite video generation emerges from autoregressive self-rollback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 40256–40265 (2026)
32. Yi, J., Jang, W., Cho, P.H., Nam, J., Yoon, H., Kim, S.: Deep forcing: Training-free long video generation with deep sink and participative compression. arXiv preprint arXiv:2512.05081 (2025)
33. Yin, S., Wu, C., Yang, H., Wang, J., Wang, X., Ni, M., Yang, Z., Li, L., Liu, S., Yang, F., et al.: Nuwa-xl: Diffusion over diffusion for extremely long video generation. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1309–1320 (2023)
34. 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)
35. Zhang, K., Tang, Z., Hu, X., Pan, X., Guo, X., Liu, Y., Huang, J., Yuan, L., Zhang, Q., Long, X.X., et al.: Epona: Autoregressive diffusion world model for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 27220–27230 (2025)
36. Zhang, L., Cai, S., Li, M., Wetzstein, G., Agrawala, M.: Frame context packing and drift prevention in next-frame-prediction video diffusion models. *Advances in Neural Information Processing Systems* **38**, 30546–30566 (2026)
37. Zhou, 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. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12405–12414 (2025)
38. Zhu, H., Zhao, M., He, G., Su, H., Li, C., Zhu, J.: Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. arXiv preprint arXiv:2602.02214 (2026)