HIERARCHICAL DIFFUSION FOR CAMERA-CONTROLLED AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We address the task of long-video, camera-controlled generation from a single image. The dominant autoregressive paradigm suffers from two key limitations: (1) error accumulation over time, as distant frames are conditioned on previously generated outputs rather than ground truth, leading to drift; and (2) ambiguity in how to incorporate camera control. Standard solutions to both—such as self-forcing and conditioning on explicit 6-DOF poses—either incur significant computational overhead or produce drifting scene geometries over time. We propose a simple and lightweight framework that combines autogressive generation (WanAR) with hierarchical long-context denoising (FixAR). To enforce scene geometry remains consistent with input cameras, we draw from feed-forward novel-view synthesis and reformulate video generation as a warping-and-inpainting problem. To mitigate drift, we introduce a hierarchical denoising strategy in which different stages condition on different temporal histories. Specifically, we first generate frames conditioned on the immediate past, and then refine them using a second denoiser conditioned on a temporally subsampled long-range history to enforce global consistency. This fine-to-coarse scheme enables each generated frame to effectively condition on the full temporal context without incurring a prohibitive cost. Experiments on RealEstate10k and DL3DV demonstrate that WanAR+FixAR outperforms prior methods for camera-controlled long-video generation, producing results that are both temporally stable and consistent with input camera trajectories.

1 INTRODUCTION

Recent advances in video diffusion models Blattmann et al. (2023); Yang et al. (2024); Team et al. (2025); HaCohen et al. (2024) have enabled photorealistic video generation with increasing levels of control. In this work, we study the problem of long-video generation from a single image given future camera trajectories. This setting is particularly appealing for applications such as virtual scene exploration and content creation, where controllable viewpoint changes must remain 3D consistent over extended time horizons. However, generating long videos that faithfully follow camera motion while maintaining spatial and temporal coherence remains a significant challenge.

Why is this hard? Existing approaches predominantly rely on autoregressive generation, where future frames are conditioned on previously generated outputs over a sequence of rollouts. While this is effective for short sequences, such methods accumulate errors over time, leading to drift in long videos. A second challenge lies in how camera motion is incorporated. One widely adopted approach is to condition on explicit 6-DOF pose representations, such as Plücker coordinates or ray maps. One subtle challenge is that many flagship benchmarks (like RealEstate10K Zhou et al. (2018)) provide cameras only up-to-scale, making camera-conditioned video generation somewhat undefined - e.g., the world looks different when moving forward 10cm versus 10m.

Status quo. Prior work has attempted to address these challenges along two main directions. To mitigate autoregressive drift, several methods adopt self-forcing strategies Huang et al. (2025), where models are periodically trained or evaluated on their own generated outputs. While effective in reducing error accumulation, these approaches significantly increase computational cost and are difficult to scale to long sequences; there is a linear increase in time and space for generating training data as rollouts can only be sequentially obtained. Most camera-controlled video generation meth-

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

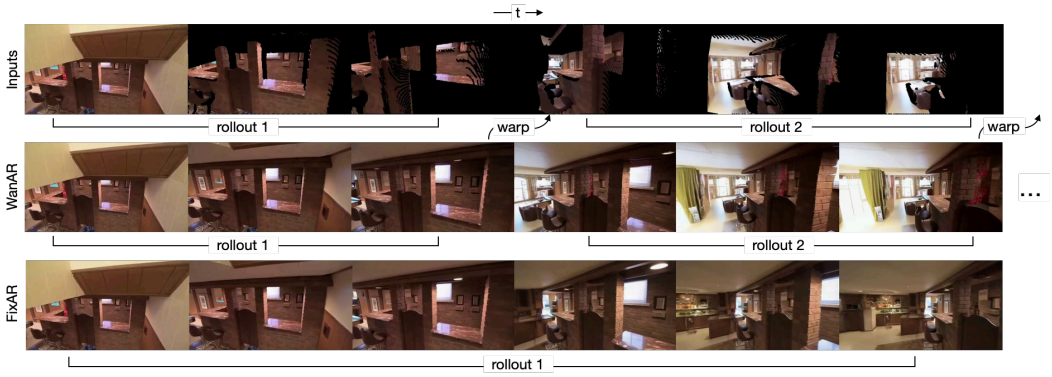

Figure 1: In this work, we consider the problem of camera-controlled video generation. Existing approaches condition on cameras represented as ray encodings or Plucker coordinates, but these learn poses up to a global scale factor. We draw from literature in feed-forward novel-view synthesis that warp given views into future poses (**top**) and reformulate video generation as video inpainting (**middle**). While this allows for precise camera control, minute-long autoregressive video generation in this setup still suffers from drift and error accumulation over time. We propose a hierarchical denoising scheme (**bottom**) that looks at multiple rollouts from the inpainting stage and corrects for long horizon artifacts.

ods appear to circumvent the up-to-scale ambiguity by overfitting to the scale of a particular dataset, making such camera-conditioned generation models unlikely to generalize.

Our approach. We address these challenges with a simple and lightweight framework that combines online reconstruction with hierarchical denoising. First, to respect the fixed but unknown scale of input camera trajectories, we draw inspiration from feed-forward novel-view synthesis and reformulate video generation as a warping-and-inpainting problem, aligning intermediate depth predictions to the scale of the ground-truth geometry (typically provided in the same scale as camera inputs, often obtained via COLMAP). Second, to mitigate autoregressive drift, we propose a hierarchical denoising strategy in which different stages condition on different temporal histories. Concretely, we generate frames in a fine-to-coarse manner: an initial denoiser conditions on the immediate past to produce high-frequency details, while a second denoiser conditions on a temporally subsampled long-range history to enforce global consistency. This design allows each generated frame to efficiently incorporate information from the entire past without incurring the cost of conditioning on an exhaustive set of historical frames.

In summary, our contributions are as follows: (1) we identify scale inconsistency as a key failure mode of pose-conditioned video generation and propose a warping-based formulation that enforces scale consistency between the scene and camera trajectory; (2) we introduce a hierarchical denoising scheme that reduces long-horizon drift by conditioning on multi-scale temporal context; and (3) we present a simple combination of models, WanAR and FixAR, that achieves state-of-the-art performance on RealEstate10k and DL3DV for camera-controlled long-video generation, producing videos that remain consistent with the scale of input camera trajectories.

2 Related Work

Autoregressive long video generation. Modern video diffusion models Ho et al. (2022); Blattmann et al. (2023); HaCohen et al. (2024); Yang et al. (2024); Team et al. (2025); Chen et al. (2025c) can generate photorealistic videos, but are typically limited to short temporal horizons. To extend generation length, autoregressive formulations are widely adopted to roll out longer sequences Chen et al. (2025a); Teng et al. (2025); Deng et al. (2024); Jin et al. (2024); Chen et al. (2024); Yin et al. (2025); Song et al. (2025); Huang et al. (2025); Cui et al. (2025). However, a fundamental challenge in autoregressive video diffusion is *drift*, which progressively degrades visual quality over long rollouts. Recent approaches attempt to mitigate this issue by either exposing models to drifted distributions during training Chen et al. (2024); Song et al. (2025); Huang et al. (2025); Zhu et al.

(2026); Liu et al. (2025), or by modifying attention mechanisms and introducing memory compression modules to reference earlier, cleaner context during inference Zhang et al. (2025b); Cai et al. (2025); Zhang et al. (2025c); Yang et al. (2025); Li et al. (2026b). While these methods improve long-range consistency, they still suffer from limitations. For example, diffusion-forcing Chen et al. (2024); Song et al. (2025) relies on approximated drift distributions and exhibits a train–test drift distribution mismatch, while self-forcing Huang et al. (2025) incurs increased training cost due to multi-step rollouts at every training iteration. In this work, we propose to decouple autoregressive generation (WanAR) and drift correction (FixAR) via a hierarchical denoising framework, where FixAR is trained on precomputed data pairs with real drift.

Camera-conditioned video world model. By introducing action controllability into long video generation and making the underlying world behind visual content explorable, video world models have emerged and demonstrated strong performance Bruce et al. (2024); Ball et al. (2025); Valevski et al. (2024); Team et al. (2026a); Hong et al. (2025); Team et al. (2026b). Most video world models primarily used keyboard inputs as a simple form of action control He et al. (2025b); Guo et al. (2025); Etched AI (2025), while many recent works Yu et al. (2025a); Xiao et al. (2025); Kong et al. (2025); Li et al. (2025b); Zong et al. (2020); Wu et al. (2025c); Gao et al. (2025) increasingly emphasizes camera motion as the control signal to enable more precise manipulation and support specific application scenarios. A natural question is how to incorporate camera information into video generative models. Existing approaches can be broadly categorized into three groups: (i) direct camera encoding, (ii) injecting 3D geometry through positional encoding fields Li et al. (2025a); Bai et al. (2025b), and (iii) injecting 3D geometry via rendering in pixel space. Most existing methods Yu et al. (2025a); Xiao et al. (2025); Shen et al. (2026a); Li et al. (2025b); Team et al. (2026b;a); Hong et al. (2025); He et al. (2025b); Hong et al. (2025); Yu et al. (2025a) fall into the first category, where normalized representations (e.g., Plücker rays or ray maps) are encoded into latent space to guide the generation process, allowing camera control to be learned from large-scale camera annotations. However, such latent-space conditioning often leads to imprecise control when reflected in pixel-space generation, and tends to generalize poorly to out-of-distribution camera trajectories. The second line of work Shen et al. (2026a); Zhang et al. (2025a) models 3D geometric transformations induced by camera motion through modifications to positional encoding. This approach achieves more accurate camera control, but the modified architecture introduces additional training overhead. In this work, we adopt the third line of work Wu et al. (2025c); Kong et al. (2025), which directly models camera control by estimating 3D geometry using a metric depth estimator and warping it into future frames in an online fashion. This approach enables accurate camera control while naturally reformulating autoregressive generation as an inpainting task.

Generative novel-view synthesis. Current state-of-the-art novel-view synthesis (NVS) methods are largely based on generative diffusion models Gao et al. (2024); Wu et al. (2025b); Watson et al. (2022); Yu et al. (2025c;b); Chen et al. (2025b); Bai et al. (2025a); He et al. (2025a); Ren et al. (2025); Li et al. (2026a); Bai et al. (2025b); Seo et al. (2024); Zhao et al. (2024). They are closely related to camera-conditioned video world models, but their spatial and temporal horizons for novel-view generation remain limited. Early NVS approaches also widely adopted direct camera encoding in latent space Gao et al. (2024); Wu et al. (2025b), but lacked precise camera control. More recent methods instead employ 3D geometry-based warping-and-inpainting formulations Yu et al. (2025c;b); Chen et al. (2025b); Ren et al. (2025), leading to improved visual quality and more accurate camera control.

In addition, there is a line of work in NVS that uses diffusion models to correct rendering artifacts caused by large viewpoint changes Wu et al. (2025a); de Lutio et al. (2026). This perspective is closely related to the motivation behind our FixAR. In our setting, we interpret autoregressive drift in long-horizon video generation as another form of accumulated artifact over time, which can be corrected using video diffusion.

3 METHOD

3.1 PRELIMINARIES: FLOW MATCHING VIDEO DIFFUSION

We adopt the open-source DiT-based Peebles & Xie (2023) video diffusion model Wan-I2V-14B Team et al. (2025) as our base model. Unlike traditional diffusion models that rely on score matching and stochastic reverse diffusion, Wan is trained with rectified flow matching Esser et al.

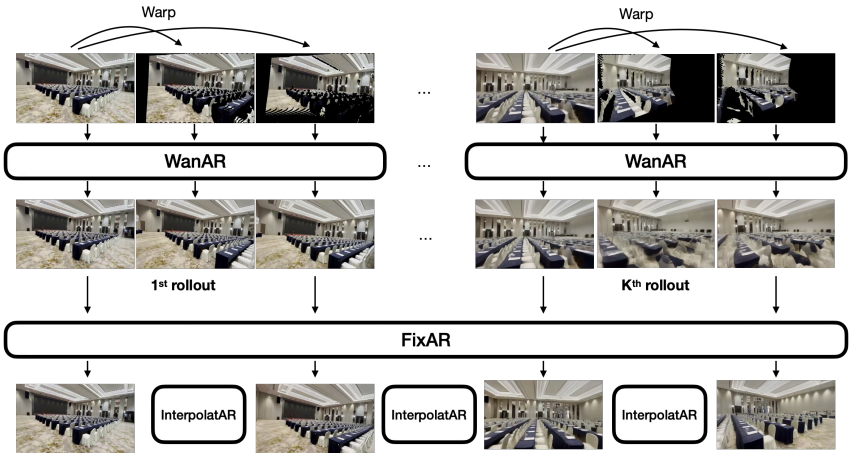

Figure 2: **Overview of WanAR + FixAR.** We build upon WAN Team et al. (2025), a video diffusion model, and adapt it for camera-controlled autoregressive video generation in static 3D scenes. At each rollout, we estimate depth using MoGe-2 Wang et al. (2025c) on the input frame (or the last generated frame from the previous rollout), and warp it into future camera poses to provide geometry-aware conditioning for WanAR (**top**). While WanAR can generate plausible long videos autoregressively, visual drift and artifacts gradually accumulate over multiple rollouts. To mitigate this issue, we introduce a hierarchical denoising framework with a second-stage denoiser, FixAR (**middle**), which refines temporally subsampled keyframes generated across different rollouts using a broader temporal context. Finally, since FixAR only operates on sparse keyframes, we apply InterpolatAR to synthesize intermediate frames between adjacent refined keyframes (**bottom**). This hierarchical fine-to-coarse formulation enables both accurate camera control and temporally stable long-horizon video generation.

(2024); Liu & Gong (2023); Lipman et al. (2023), which formulates generation as solving an ordinary differential equation (ODE) without stochastic terms. This formulation provides a stable and scalable training objective while preserving a likelihood-based interpretation.

Specifically, given a clean video latent x_1, a random noise latent $x_0 \sim \mathcal{N}(0, I)$, a conditioning signal c, and a timestep $t \in [0, 1]$, flow matching defines the interpolated latent as

$$x_t = (1 - t)x_0 + tx_1.$$

The model is trained to predict the constant velocity field that transports noise x_0 to data x_1:

$$\mathcal{L} = \|f_\theta(x_t, c, t) - (x_1 - x_0)\|_2^2 .$$

3.2 WanAR: Camera control for autoregressive video generation

We consider the problem of camera-conditioned long video generation: given an input image I_0 and a target camera trajectory $\{C_t\}_{t=0}^T$, the goal is to generate future observations $\{I_t\}_{t=1}^T$ that follow the specified camera motion, where T can be large. General-purpose video diffusion models such as Wan-I2V-14B cannot be directly applied to this setting, because they (i) do not explicitly support camera conditioning, and (ii) are limited to a short generation horizon L, where typically $L \ll T$. To address this, we build **WanAR** on top of Wan. WanAR achieves camera control by propagating explicit 3D geometry (depth and camera), and extends generation length through an autoregressive formulation.

Given the initial observation I_0, we first estimate its metric depth D_0 using the monocular depth estimator MoGe-2 Wang et al. (2025b;c). Since the corresponding camera C_0 is known, this gives us sufficient 3D geometry for the first frame. We then propagate this information to future viewpoints within the base model horizon by forward warping:

$$(\tilde{I}_t, M_t) = \mathcal{W}(I_0, D_0, C_0, C_t), \qquad t = 1, \ldots, L - 1,$$

where $\tilde{I}_t$ denotes the warped partial observation at time t, M_t is the corresponding inpainting mask, and $\mathcal{W}$ is the warping function.

Although the warped frames contain only partial RGB observations, they already provide accurate camera-conditioned spatial guidance. As a result, camera-conditioned video generation is reformulated as an inpainting problem that completes these partial observations:

$$\{\hat{I}_t\}_{t=1}^{L-1} = f_\theta\Big(I_0, \{(\tilde{I}_t, M_t)\}_{t=1}^{L-1}\Big),$$

where $\hat{I}_t$ denotes the generated full frame. To adapt Wan-I2V to this task, we modify its conditioning mechanism. In the original I2V model, the future frames and masks are represented by zero-padded latents in VAE space. In contrast, we replace this zero padding with the actual latent encodings of the warped frames and their corresponding inpainting masks.

So far, this formulation enables camera control only within the base diffusion horizon. To extend it beyond length L, we apply the same inpainting process autoregressively. At rollout k, we take the latest anchor frame $\hat{I}_{k(L-1)}$, estimate its depth online using MoGe-2 Wang et al. (2025b;c), and warp it to the next chunk of target cameras. Formally, for

$$k = 0, \ldots, \left\lceil \frac{T}{L-1} \right\rceil - 1,$$

we generate

$$\{\hat{I}_t\}_{t=k(L-1)+1}^{(k+1)(L-1)} = f_\theta\Big(\hat{I}_{k(L-1)}, \{(\tilde{I}_t, M_t)\}_{t=k(L-1)+1}^{(k+1)(L-1)}\Big).$$

In this way, WanAR produces a long video sequence that follows the entire target camera trajectory.

3.3 FixAR: Hierarchical denoising for mitigating drift

The autoregressive formulation of WanAR enables long-horizon generation, but it also introduces drift: except for the first rollout, each new chunk is conditioned on previously synthesized frames, so errors accumulate over time and gradually degrade visual quality. To mitigate this effect, we introduce **FixAR**, a second-stage fixing module that is applied after every few WanAR rollouts.

Specifically, after every K rollouts, we uniformly sample a sparse set of keyframes from the generated sequence and feed them into FixAR for content correction. Let $\{\hat{I}_{\tau_j}\}_{j=0}^{L-1}$ denote the selected sparse keyframes. FixAR outputs the corrected frames

$$\{I^\star_{\tau_j}\}_{j=0}^{L-1} = g_\psi\Big(\{\hat{I}_{\tau_j}\}_{j=0}^{L-1}\Big),$$

where g_ψ denotes the FixAR model and $I^\star_{\tau_j}$ are the corrected keyframes.

To train FixAR on realistic drift patterns, we construct its training set by following the same WanAR rollout process and the same sparse keyframe sampling strategy used at inference time. Unlike self-forcing approaches Huang et al. (2025), FixAR training does not require multi-step inference during each training iteration; instead, it uses precomputed input-target pairs, which substantially reduces computational overhead. Once trained, FixAR can be applied either globally to an entire generated sequence or autoregressively at a low frequency.

FixAR mitigates drift for two reasons. First, due to its hierarchical design, it can directly reference earlier frames with higher visual quality, or even ground-truth frames during training, allowing appearance and style to be propagated over a longer horizon. Second, because it is trained on realistic drift distributions, it learns to identify artifact-prone regions and correct them accordingly.

Recovering the original video framerate. Because of the limited base model horizon and for computational efficiency, FixAR operates only on sparse keyframes. To recover the full video sequence at the original framerate, we train an additional video diffusion model, **InterpolatAR**, to interpolate between neighboring corrected keyframes. This is a substantially easier task, since adjacent corrected keyframes are already visually and temporally close, and little hallucination is required. Importantly, InterpolatAR does not introduce additional drift, because it only fills in intermediate content between already generated frames, rather than performing autoregressive extrapolation as in WanAR.

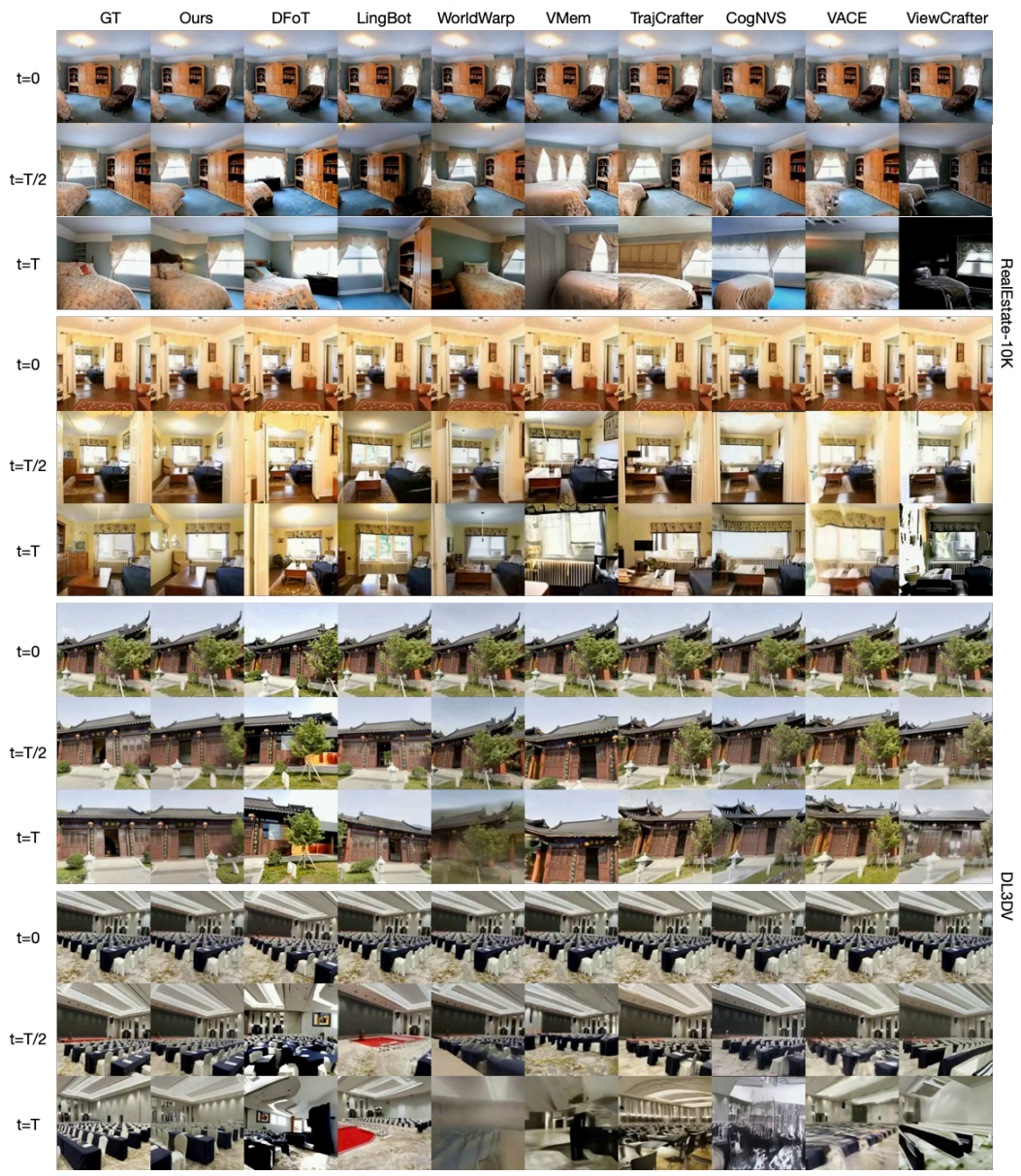

Figure 3: **Qualitative comparison on RealEstate10k and DL3DV.** We show qualitative results on two test scene from RealEstate10k (top two examples) and DL3DV (bottom two examples) across a range of video generation, and novel-view synthesis baselines. We see that proposed WanAR + FixAR combination of models "renders" the scene using scale-consistent camera poses (as compared to ray encoding in DFoT or Plucker conditioning in LingBot), and show that hierarchical denoising mitigates drift over time which some autoregressive baselines like VMem struggle with. We also note that TrajectoryCrafter, a method designed for novel-view synthesis (in short sequences) is also surprisingly strong at the task of camera-controlled minute-long video generation.

4 Empirical Analysis

4.1 Experimental setup

Datasets. We train and evaluate our model on two real-world datasets, RealEstate10K Zhou et al. (2018) and DL3DV Ling et al. (2024), which cover diverse indoor and outdoor scenes. RealEstate10K Zhou et al. (2018) contains approximately 60K videos, with ground-truth camera

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓	mPSNR ↑
ViewCrafter Yu et al. (2025c)	12.45	0.361	0.499	67.36	0.0267	595.25	13.44
TrajCrafter Yu et al. (2025b)	14.35	0.428	0.410	30.54	0.0042	347.75	15.54
CogNVS Chen et al. (2025b)	13.22	0.379	0.474	48.58	0.0116	457.35	14.39
VACE Jiang et al. (2025)	15.23	0.487	0.426	49.62	0.0119	393.30	16.89
Lingbot-World Team et al. (2026b)	12.90	0.362	0.513	25.46	**0.0012**	398.31	13.36
VMem Li et al. (2025b)	12.13	0.336	0.583	49.12	0.0099	465.13	12.46
DFoT Song et al. (2025)	13.95	0.423	0.422	38.28	0.0084	387.21	15.17
WorldWarp Kong et al. (2025)	13.81	0.425	0.475	38.59	0.0042	451.01	14.55
WanAR + FixAR	**16.69**	**0.525**	**0.341**	**23.16**	0.0018	**282.59**	**18.31**

Table 1: **Comparison to state-of-the-art on RealEstate10K Zhou et al. (2018).** We compare our hierarchical diffusion strategy to state-of-the-art approaches for long video generation and novel-view synthesis of static scenes. Our first finding is that a combination of WanAR + FixAR beats all prior art on RealEstate10K, which we attribute to better camera control in WanAR, and better drift mitigation from FixAR. One surprising finding is that novel-view synthesis approaches like TrajectoryCrafter in fact perform better at this task than many long video generation works. With mPSNR, we show better retention of visible regions particularly driven by precise camera control.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓	mPSNR ↑
ViewCrafter Yu et al. (2025c)	13.55	0.307	0.459	74.76	0.0114	847.25	13.19
TrajCrafter Yu et al. (2025b)	13.49	0.306	0.428	53.45	0.0042	703.82	13.93
CogNVS Chen et al. (2025b)	12.70	0.268	0.467	81.09	0.0166	882.71	13.04
VACE Jiang et al. (2025)	13.26	0.302	0.476	92.48	0.0209	781.41	13.77
Lingbot-World Team et al. (2026b)	12.87	0.246	0.457	**40.26**	**0.0013**	742.09	13.07
VMem Li et al. (2025b)	12.02	0.232	0.563	80.76	0.0138	867.67	12.13
DFoT Song et al. (2025)	9.19	0.143	0.603	82.44	0.0232	931.15	9.23
WorldWarp Kong et al. (2025)	12.63	0.265	0.540	74.81	0.0072	1002.4	12.73
WanAR + FixAR	**13.84**	**0.320**	**0.421**	47.75	0.0034	**640.64**	**14.05**

Table 2: **Comparison to state-of-the-art on DL3DV Ling et al. (2024).** We compare our method with prior art on the DL3DV dataset which has more challenging camera motion, like sudden translation changes, and 360° camera pan. Despite this, our method outperforms baselines by a considerable margin. Once again, mPSNR demonstrates better camera control with our warp-and-inpaint strategy for long video generation.

annotations available for both training and testing. For DL3DV Ling et al. (2024), since the released version does not provide camera annotations, we use a COLMAP-preprocessed training set Wang et al. (2025a) with approximately 7K videos. For the test set, we process the videos with MegaSAM Li et al. (2025c) to obtain the corresponding camera poses and depth maps.

Baselines. We compare our method with baselines from both camera-conditioned long video generation and novel-view synthesis. For camera-conditioned video generation, we compare with VMem Li et al. (2025b), DFoT Song et al. (2025), Lingbot-World Team et al. (2026b), and WorldWarp Kong et al. (2025). VMem, DFoT, and Lingbot-World use Plücker rays or ray maps as camera-control representations, while WorldWarp uses 3D Gaussian rendering as camera guidance. Lingbot-World is an open-source industrial model, and we use its flagship bidirectional model for evaluation.

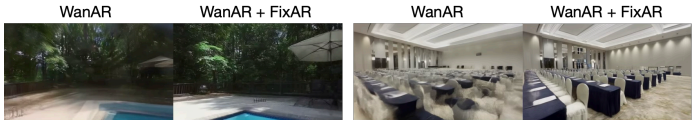

Figure 4: **Qualitative analysis of hierarchical denoising.** WanAR exhibits drift over long rollouts, while FixAR mitigates the drift and cleans up accumulated artifacts.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓
WanAR	16.63	0.519	0.342	36.68	0.0022	455.87
WanAR + FixAR	16.71	0.521	0.338	34.55	0.0018	429.14

Table 3: **Quantitative analysis of hierarchical denoising.** FixAR improves generation realism and leads to consistent gains across evaluation metrics.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓	mPSNR ↑
Warp-and-inpaint	16.63	0.519	0.342	36.68	0.0022	455.87	18.20
Plucker ray	14.95	0.438	0.416	35.62	0.0021	505.49	15.33

Table 4: **Comparison of different conditioning strategies.** The Plucker baseline adds a convolutional adapter to the Wan2.1-14B backbone, and is finetuned on Re10K using the same setup.

For novel-view synthesis, we adopt video diffusion-based methods that also use warp-and-inpaint-style camera control, including ViewCrafter Yu et al. (2025c), TrajCrafter Yu et al. (2025b), and CogNVS Chen et al. (2025b). We further include VACE Jiang et al. (2025), a Wan-14B-based video inpainter, as an additional baseline. Since these methods are designed for fixed-length rendering, we evaluate them in an image-to-video autoregressive manner, following the same rollout setting as our method.

Metrics. We follow standard evaluation protocols used in related work Xiao et al. (2025); Kong et al. (2025); Li et al. (2025b); Shen et al. (2026b), including pairwise reconstruction metrics PSNR, SSIM, and LPIPS, as well as distribution-level metrics FID and KID. We also report FVD to evaluate video generation quality. In addition, we adopt masked PSNR (mPSNR) Gao et al. (2022), which computes PSNR only over the covisible region from the first frame. This metric better reflects camera-control accuracy without scale ambiguity.

Since some baselines Song et al. (2025); Li et al. (2025b) only support a resolution of 256×256 and a temporal frame skip of 10, we crop and resize the outputs of all methods, including ours, to the same spatial and temporal resolution for fair comparison.

Implementation details. We initialize our models from the official Wan-I2V-14B-480P checkpoint Team et al. (2025) and fine-tune them using LoRA with rank 32. We use a batch size of 8 and a learning rate of 1×10^{-4}. For each model, including WanAR, FixAR, and InterpolatAR, we fine-tune for 10K steps on each dataset using 8 H100 GPUs at a resolution of 256×256. Training each model takes approximately one day. Further inference details are provided in Appendix .1.

4.2 Comparison to state-of-the-art

RealEstate10K. Tab. 1 presents a detailed comparison between existing methods and ours on RealEstate10K. Overall, NVS methods achieve stronger reconstruction metrics, including PSNR, SSIM, LPIPS, and mPSNR, mainly because their warp-and-inpaint formulation provides more accurate camera control. In contrast, camera-conditioned long video generation methods perform better on visual fidelity metrics, such as FID, KID, and FVD, since they are typically trained with dedicated anti-drift strategies and are therefore better at preserving long-term visual quality. The importance of accurate camera control is also illustrated in Fig. 3. For example, Lingbot-World, despite being an industrial-level bidirectional video diffusion model capable of generating highly photorealistic videos, still obtains relatively poor PSNR due to inaccurate camera control. This suggests that Plücker-ray-based conditioning is not sufficiently accurate or generalizable for this setting. In contrast, our method achieves state-of-the-art performance across major metrics by combining accurate 3D-informed warp-and-inpaint camera control with hierarchical denoising for drift mitigation.

DL3DV is a more challenging dataset, containing fast viewpoint transitions, extreme camera trajectories, and diverse scene content. As shown in Fig. 3, many NVS methods suffer from distorted content, largely because depth estimation becomes unreliable on drifted generated frames during autoregressive rollout. In contrast, our method uses FixAR to suppress drift accumulation, ensuring that the corrected frames remain meaningful inputs for subsequent depth estimation and warping. As reported in Tab. 2, our method again achieves state-of-the-art overall performance, despite marginally lower FID and KID compared with Lingbot-World. These results demonstrate that our method generalizes well to diverse indoor and outdoor scenarios with challenging camera motion.

4.3 Ablations and analysis

Effect of hierarchical denoising. We conduct ablation studies on a 50-sequence subset of RealEstate-10K. Fig. 4 and Tab. 3 show the qualitative and quantitative effects of the hierarchical

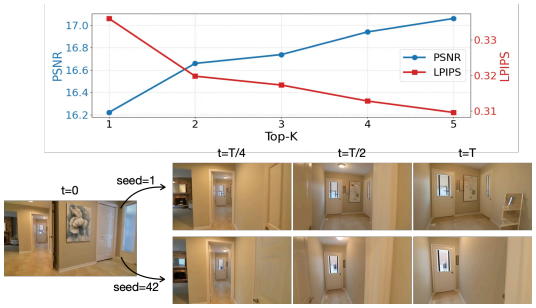

Figure 5: **Top-k evaluation.** We note that generative methods for camera-controlled video generation can produce multi-modal outputs for unseen regions of the scene as camera pans in new direction. We conduct such an evaluation for our method. Our finding is that the performance of our final predictions improves with respect to the ground-truth as more samples are drawn from WanAR (thereby covering the generation of multiple plausible 3D scenes in the future).

denoiser. By mitigating drift, the denoiser improves generation realism and consistently improves the evaluation metrics.

Ablations on camera pose conditioning. We compare WanAR's warp-and-inpaint conditioning with Plucker ray conditioning using the same backbone and training setup. As shown in Tab. 4, warp-and-inpaint improves reconstruction quality, camera-control accuracy, and video quality, while achieving comparable FID and KID. This suggests that warp-and-inpaint provides more precise camera control in our setting. Additional ablations are reported in Appendix .3.

Evaluation of multi-modal generation. A key advantage of video diffusion models is their multi-modal generation capability, which we also observe in WanAR + FixAR. As shown in Fig. 5, the same corridor scene can be plausibly generated as either a narrow straight pathway or a larger room space connected to another room by an additional door. Both outcomes are reasonable under the given input. We therefore evaluate Top-K generation metrics by sampling multiple outputs with different random seeds.

5 Discussion

We present a simple framework for long-video, camera-controlled generation that combines autoregressive generation with hierarchical denoising. Our results show that reformulating generation as a warping-and-inpainting problem helps maintain consistency with input camera trajectories, particularly in the presence of up-to-scale ambiguity, while hierarchical denoising improves temporal stability by allowing each frame to incorporate both short- and long-range temporal information. Compared to prior approaches that rely on self-forcing or explicit pose conditioning alone, our method achieves stronger long-horizon consistency. These findings suggest that both scale consistency and long-range temporal conditioning are important for generating stable, camera-controlled videos over extended sequences.

Limitations. We note that our proposed warp-and-inpaint hierarchical video generation can only be used for static scenes. WanAR relies on the quality of underlying geometry, and errors in depth or reconstruction can propagate through the warping process, leading to artifacts or inconsistencies. WanAR rollouts only warp the last frame of the previous rollout and therefore do not have any memory mechanism baked in. This means that loop closure in long sequences may not be accurate. While hierarchical denoising with FixAR reduces drift, it does not fully eliminate it as our second stage operates over a finite horizon of frames too. Finally, our models trained on one dataset cannot be used for another dataset if the scale distributions between the two datasets are drastically different. We cover a detailed inference cost of our method with respect to other baselines.

Broader impacts. Efficient and temporally consistent camera-controlled video synthesis could broaden access to tools for creative media production, 3D exploration in virtual reality applications, and embodied AI research, though care should be taken to mitigate risks associated with realistic synthetic video generation.

AI USE STATEMENT

In this work, we used generative AI tools for [design or provide feedback on research methodology or experiments; implement methods]. We have not used generative AI tools for [help develop theoretical models or conceptual frameworks; formulate mathematical claims; provide critical ingredients for proving mathematical claims; assist in the writing of proofs; propose or refine hypotheses; assist with translation; support qualitative and thematic data analysis; interpret results], and [Generate synthetic data sets; clean and reformat dataset] are not applicable to this work. Additionally, we used generative AI tools for [edit a research paper to improve readability; summarize or analyse existing literature; create or edit software code;]. We have reviewed all AI-assisted work. We checked AI-assisted revisions so it closely follows what we propose in the paper and does not make factual mistakes. We verified and tested the code generated by LLM. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To make the results reproducible, we provide the key design choices and implementation details, including training and inference procedures, in Sections 4.1 and .1. We will release our code after internal review.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, 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, 2025a.

Yunpeng Bai, Haoxiang Li, and Qixing Huang. Positional encoding field. *arXiv preprint arXiv:2510.20385*, 2025b.

Philip J. Ball, Jakob Bauer, Frank Belletti, et al. Genie 3: A new frontier for world models. `https://deepmind.google/models/genie/`, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024.

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

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025a.

Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. In *Advances in Neural Information Processing Systems*, 2025b.

Kaihua Chen, Deva Ramanan, and Tarasha Khurana. Using diffusion priors for video amodal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22890–22900, 2025c.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. In *International Conference on Learning Representations*, 2025.

Riccardo de Lutio, Tobias Fischer, Yen-Yu Chang, Yuxuan Zhang, Jay Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Katarina Tothova, Zan Gojcic, and Haithem Turki. Artifixer: Enhancing and extending 3d reconstruction with auto-regressive diffusion models. *arXiv preprint arXiv:2603.00492*, 2026.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *International Conference on Learning Representations*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Etched AI. Open-oasis. `https://github.com/etched-ai/open-oasis`, 2025.

Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems*, volume 35, pp. 33768–33780, 2022.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Junhao Zhuang, Chengming Xu, Jianfeng Feng, Yu Qiao, Yanwei Fu, Chenyang Si, and Ziwei Liu. Longvie 2: Multimodal controllable ultra-long video world model. *arXiv preprint arXiv:2512.13604*, 2025.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T. Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: A real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. 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, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 8633–8646, 2022.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, 2025.

Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. Vace: All-in-one video creation and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17191–17202, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *International Conference on Learning Representations*, 2024.

Hanyang Kong, Xingyi Yang, Xiaoxu Zheng, and Xinchao Wang. Worldwarp: Propagating 3d geometry with asynchronous video diffusion. *arXiv preprint arXiv:2512.19678*, 2025.

Nathanael Kuo, Homer Li, Cheng Peng, and Myron Brown. Wriva public data (ultrra challenge), 2025. URL https://ieee-dataport.org/open-access/wriva-public-data.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. Rerope: Repurposing rope for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026a.

Haodong Li, Shaoteng Liu, Zhe Lin, and Manmohan Chandraker. Rolling sink: Bridging limited-horizon training and open-ended testing in autoregressive video diffusion. *arXiv preprint arXiv:2602.07775*, 2026b.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. In *Advances in Neural Information Processing Systems*, 2025a.

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

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025c.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, 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.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Xingchao Liu and Chengyue Gong. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6121–6132, 2025.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. In *Advances in Neural Information Processing Systems*, volume 37, pp. 80220–80243, 2024.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026a.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, et al. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026b.

Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. In *International Conference on Machine Learning*, 2025.

HY-World Team, Chenjie Cao, Xuhui Zuo, Zhenwei Wang, Yisu Zhang, Junta Wu, Zhenyang Liu, et al. Hy-world 2.0: A multi-modal world model for reconstructing, generating, and simulating 3d worlds. *arXiv preprint arXiv:2604.14268*, 2026a.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026b.

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

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, 2024.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025a.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5261–5271, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, 2025c.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. Difix3d+: Improving 3d reconstructions with single-step diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26024–26035, 2025a.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T. Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26068, 2025b.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. In *Advances in Neural Information Processing Systems*, 2025c.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. In *Advances in Neural Information Processing Systems*, 2025.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

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

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 100–111, 2025b.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Cruz Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. *arXiv preprint arXiv:2512.07237*, 2025a.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, 2025b.

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

Yizhou Zhao, Hengwei Bian, Kaihua Chen, Pengliang Ji, Liao Qu, Shao-yu Lin, Weichen Yu, et al. Metric from human: Zero-shot monocular metric depth estimation via test-time adaptation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 104724–104753, 2024.

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

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

Jonathan Zong, Dhiraj Barnwal, Rupayan Neogy, and Arvind Satyanarayan. Lyra 2: Designing interactive visualizations by demonstration. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):304–314, 2020.

APPENDIX

In this supplement, we include further implementation details of implementation (Section .1), computational cost (.2), and more ablations .3 of our WanAR + FixAR approach. We further present a nuanced discussion of camera control in video generation (Section .4).

.1 MORE IMPLEMENTATION DETAILS

Given one initial image and a sequence of L target cameras, our objective is to generate an L-frame video following the target camera trajectory. WanAR, FixAR, and InterpolatAR all use the same Wan2.1-14B architecture, but have separately trained checkpoints for their respective tasks. Unless otherwise specified, each stage operates at a resolution of 336×592 with a temporal window of T=49 frames.

For InterpolatAR, we place the corrected FixAR keyframes at their corresponding temporal locations and replace the missing intermediate frames with blank frames. The first frame of each input window is always an observed keyframe. InterpolatAR then predicts the missing frames conditioned on the visible keyframes (i.e. if FixAR was applied at a keyframe interval of 4 then frames 1, 5, 9 ... are input to InterpolatAR which interpolates to get frames 2, 3, 4, 6, 7, 8 ...).

For RealEstate10K, we generate (L=280) frames at 10 FPS: 1) WanAR: We subsample the target trajectory using an interval of 2 to generate the sequence with 3 autoregressive diffusion calls. 2) FixAR: We select 49 generated keyframes with an average temporal interval of approximately 6 and jointly refine them using 1 diffusion call. 3) InterpolatAR: We use an interval of 1 to reconstruct the complete 280-frame sequence, requiring 6 diffusion calls. The complete RealEstate10K pipeline uses 10 diffusion calls: 3 WanAR, 1 FixAR, and 6 InterpolatAR calls.

For DL3DV, we generate (L=600) frames at 30 FPS: 1) WanAR and FixAR: We subsample the target trajectory using an interval of 3 to generate the first 10s of the sequence with 2 WanAR diffusion calls. We then select 49 keyframes with an average temporal interval of 6 and jointly refine them using 1 FixAR diffusion call. Starting from the refined result, we repeat the same process with another 2 WanAR calls followed by 1 FixAR call to generate the remaining sequence. 2) InterpolatAR: We use an interval of 1 to reconstruct the complete 600-frame sequence, requiring 12 diffusion calls. The complete DL3DV pipeline uses 18 diffusion calls: 4 WanAR, 2 FixAR, and 12 InterpolatAR calls.

.2 ANALYSIS OF COMPUTATIONAL COST

We compare the computational cost of our method with the baselines. Specifically, we run inference on 280-frame RealEstate10K videos using one 80GB H100 GPU and report the average runtime, peak GPU memory, number of diffusion calls, and denoising steps per call. As shown in Tab. 5, our method has moderate runtime and GPU memory usage. We additionally report our method at $480 \times 832 \times 49$ only for a resolution-matched cost comparison; all experiments in the paper use a training resolution of $336 \times 592 \times 49$.

Although our system contains three stages, all stages reuse the same Wan2.1-14B architecture and the same input shape of $336 \times 592 \times 49$. We sequentially reload the corresponding LoRA checkpoint for each stage without introducing additional architectural modules. This unified design reduces engineering complexity. Moreover, our hierarchical design applies FixAR at a low frequency before final interpolation, limiting the total number of diffusion calls.

.3 ADDITIONAL ABLATIONS

Sensitivity of FixAR's keyframe interval. Our default FixAR processes 49 frames, is trained with keyframe intervals ranging from 2 to 6, and uses an interval of 6 during inference. We additionally train a 25-frame FixAR with keyframe intervals ranging from 4 to 12 to study the effects of input length and larger keyframe intervals. As shown in Tab. 6, the 49-frame FixAR performs consistently across intervals from 2 to 6, indicating that it is not sensitive to the precise interval within its training range. In contrast, the performance of the 25-frame FixAR decreases when the interval increases from 4 to 8 or 12. With very sparse keyframes, the temporal correspondence between

Method	Backbone	Resolution	GPU Memory	Run Time	# Diffusion Calls
ViewCrafter	StableVideoDiffusion-2B	$576 \times 1024 \times 25$	22GB	15min	12
TrajCrafter	CogVideoX-5B	$480 \times 720 \times 49$	48GB	18min	6
CogNVS	CogVideoX-5B	$480 \times 720 \times 49$	23GB	16min	6
VACE	Wan2.1-14B	$480 \times 832 \times 49$	51GB	19min	6
VMem	SEVA-1.3B	$576 \times 576 \times 8$	6GB	4min	6
DFoT	DiT-XL/2-0.68B	$256 \times 256 \times 8$	7GB	5min	44
WorldWarp	Wan2.1-1.3B	$480 \times 832 \times 81$	36GB	8min	4
Lingbot-World	Wan2.2-14B	$480 \times 832 \times 281$	75GB	38min	1
Ours	Wan2.1-14B	$336 \times 592 \times 49$	41GB	12min	10
	Wan2.1-14B	$480 \times 832 \times 49$	48GB	17min	10

Table 5: **Computational cost comparison with baselines.** Runtime and peak GPU memory are measured on 280-frame RealEstate10K videos using a single 80GB H100 GPU. We also report our method at $480 \times 832 \times 49$ for a resolution-matched comparison.

Setting	Keyframe Interval	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓
49-length FixAR	2	16.38	0.508	0.336	34.33	0.0021	424.94
	4	16.78	0.522	0.333	34.06	0.0017	428.02
	6	16.71	0.521	0.338	34.55	0.0018	429.14
25-length FixAR	4	16.11	0.490	0.340	38.02	0.0021	461.49
	8	15.91	0.487	0.351	38.72	0.0022	488.46
	12	15.46	0.471	0.384	40.07	0.0030	561.53

Table 6: **Ablation on FixAR keyframe interval and sequence length.** We compare 49-frame and 25-frame FixAR variants under different keyframe intervals. The 49-frame model remains stable across its training range, while larger intervals degrade the performance of the 25-frame model.

adjacent input frames becomes weaker, making refinement more difficult. The 49-frame FixAR also consistently outperforms the 25-frame variant; a longer input sequence can cover a longer temporal range while retaining relatively dense keyframes.

Effect of different depth estimators. We estimate the depth of the first frame using each method and then perform one WanAR rollout. As shown in Tab. 7, the performance differences among the three depth estimators is relatively small, indicating that WanAR is robust to the choice of depth estimator – it can recover from errors introduced by depth estimators. Depth Anything V3 and MegaSAM depth provide slightly better reconstruction and video-quality metrics.

.4 Extended discussion on camera-controlled video generation

In this section, we elucidate our claim that camera-controlled video generation is a special case of novel-view synthesis (NVS) and discuss relevant paradigms for evaluation which are inspired by the literature in NVS.

Case I: Image-to-Video generation with pose / geometry information

When input observations are accompanied by reconstruction information, such as calibrated camera poses and/or 3D geometry defined in a fixed coordinate system, future camera trajectories should also be specified in that same coordinate system. This corresponds to the standard NeRF-style setting, where methods are conditioned on input images together with calibrated camera poses. More generally, this setup may also be instantiated using a single image paired with depth or XYZ information. For example, given an estimate of scene geometry, future camera motion can be defined relative to the reconstructed scene structure (e.g., moving a fixed distance toward an observed object). Framing camera control in this coordinate-aware manner removes the need to reference future ground-truth geometry during alignment and provides a cleaner formulation of camera-conditioned video generation.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓	KID ↓	FVD ↓
MoGe2	20.22	0.654	0.177	21.90	0.0013	130.58
Depth Anything v3	20.46	0.666	0.172	21.77	0.0014	129.02
MegaSAM	20.74	0.679	0.170	21.92	0.0015	125.77

Table 7: **Comparison of different geometry estimation methods.** We initialize WanAR using the first-frame depth estimated by each method and report performance after one rollout. Results are similar across estimators, indicating limited sensitivity to the choice of depth estimator.

CASE II: IMAGE-TO-VIDEO GENERATION WITHOUT POSE INFORMATION

In contrast, when only input images are provided without associated poses or geometric information of the input (such as the current setup of camera-controlled video generation works including ours), the problem more closely resembles unposed novel-view synthesis. In this setting, one possible approach is to require future camera trajectories to be specified in a metric coordinate system to avoid key pitfalls of prior art (c.f. Figure 6). An alternative is to follow protocols of recent benchmarks Kuo et al. (2025) under whose formulation: (i) methods reconstruct the input observations within a locally consistent coordinate system; (ii) the benchmark provides a similarity transform aligning the reconstruction to the ground-truth frame; and (iii) future camera queries are specified within the local reconstructed coordinate system. Our current training, inference and benchmarking protocol is based on this formulation.

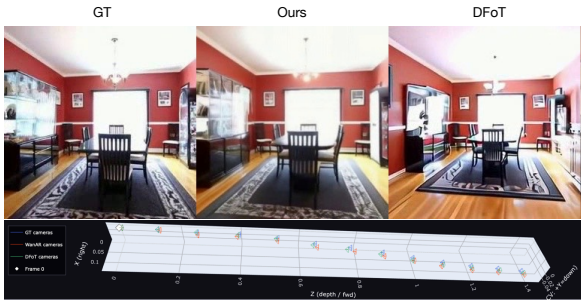

Figure 6: We identify a key limitation of existing camera-conditioned methods. When explicit 6-DOF poses are used as conditioning—e.g., via Plücker coordinates or ray maps Song et al. (2025); Team et al. (2026b)—it is difficult to enforce a consistent scale between the scene and the camera trajectory. As a result, generated videos may exhibit visual inconsistencies (**top**), reflecting scenes rendered under an arbitrary pose scale. Importantly, standard camera-control metrics Xiao et al. (2025) perform a *scale alignment* with ground-truth camera poses prior to evaluation. Consequently, even when predicted camera trajectories are incorrect up to a global scale factor, they can still achieve low error (**bottom**), masking this failure mode. Current methods evaluating performance with the PSNR metric circumvent this issue, as they overfit to the pose distribution in a particular datasets, which indicates their limited generalization to other scene or camera trajectories. In our approach, we mitigate this issue by aligning MoGE depth predictions to ground-truth depth for each rollout, and by adopting a warping-and-inpainting formulation for long-video generation. We also adopt novel-view synthesis protocols for evaluation, and instead benchmark methods for camera-control with the masked PSNR (mPSNR) metric.