WORLDNOISE: MAKING VIDEO DIFFUSION MODELS SEE THE 3D WORLD THROUGH GEOMETRIC NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in camera-controlled video diffusion models have enabled interactive world modeling, where an initial noise is progressively transformed into 3D scenes that can be explored along arbitrary camera trajectories. Yet, surprisingly, the source of these 3D worlds—randomly sampled diffusion noise—is agnostic to scene geometry. This raises a fundamental question: *can the diffusion noise itself serve as a structured representation of the 3D world?* In this paper, we introduce WORLDNOISE, a novel framework that brings 3D structure directly into the noise space of video diffusion models. Our key insight is to reinterpret diffusion noise not merely as unstructured randomness, but as a latent representation of the world that can be structured and transformed according to 3D geometry. Specifically, WORLDNOISE first constructs a 3D noise space over the observable world region jointly covered by the camera frusta along the trajectory and distributes randomness throughout this space. Then, it projects the world-space randomness into view-specific 2D noise through a carefully designed whitening and covariance-compensation process that recovers the native spatial noise prior expected by the video diffusion model. Different views therefore become different observations of the same underlying noise world, rather than being generated from unrelated noise. The resulting WORLDNOISE can be used through the existing diffusion noise interface without modifying the denoiser architecture. We further extend WORLDNOISE to long-horizon exploration by allowing the noise world to persist and grow as the camera moves through the scene. Extensive experiments demonstrate the effectiveness of WORLDNOISE, revealing a previously overlooked role of diffusion noise as a latent carrier of 3D structure and opening a new direction for 3D-aware generative world models.

1 INTRODUCTION

Camera-controlled video generation provides a generative interface to navigable 3D worlds, translating camera trajectories into visual experiences. This capability connects video generation with interactive world modeling, providing a foundation for virtual reality and embodied AI, where immersive exploration and interaction depend on a coherent spatial environment (Yu et al., 2024; Parker-Holder et al., 2024).

Recent works on camera-controlled video generation can be broadly categorized into three groups. Explicit visual conditioning renders reconstructed scene geometry along target camera trajectories to provide appearance guidance (Yu et al., 2024). Learned camera-feature conditioning encodes camera extrinsics or Plücker rays into features that guide the denoiser (Wang et al., 2024; He et al., 2025). Camera-aware positional encoding incorporates viewing geometry into the coordinates governing token interactions through relative ray and absolute orientation encoding (Zhang et al., 2026; Li et al., 2026; Yin et al., 2026). These approaches represent camera and world structure through additional learnable components or specialized architectural designs. However, they overlook the defining principle of diffusion models: transforming random noise into structured visual patterns. They leave the stochastic starting point of generation independent of scene geometry. This raises a fundamental question: can the diffusion noise itself serve as a structured representation of the 3D world?

In this paper, we explore the potential of structured noise as a latent representation. Existing studies preliminarily explore structured noise in video diffusion. FreeNoise (Qiu et al., 2024) reschedules

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.

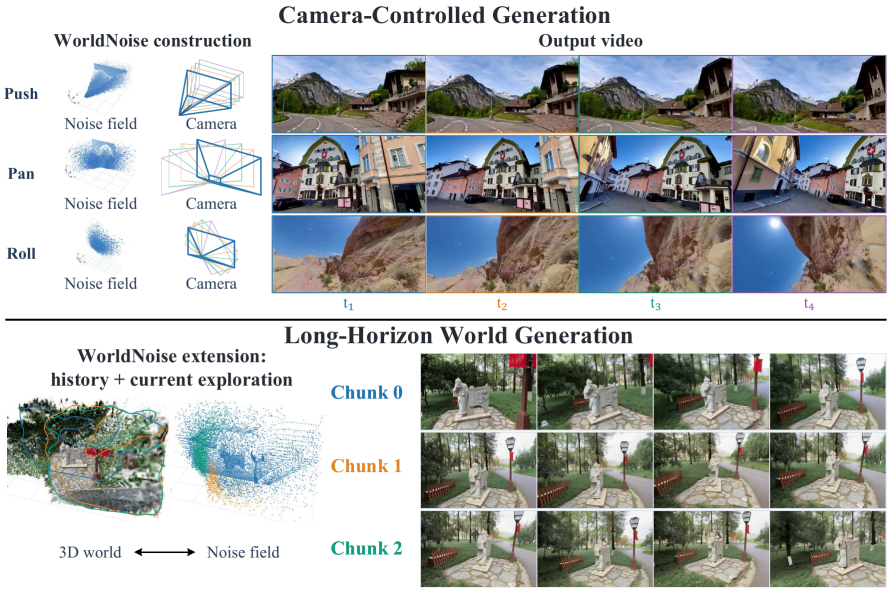

Figure 1: **WORLDNOISE**: Generating consistent views from a shared 3D noise world.

noise to establish long-range temporal correlations. FreeInit (Wu et al., 2024) iteratively refines the low-frequency spatiotemporal components of the initial noise to improve appearance and temporal consistency. Go-with-the-Flow (Burgert et al., 2025) warps noise according to optical flow to control object and camera motion. These findings show the potential of noise to carry spatial patterns and their evolution over time. Motivated by these works, we propose WORLDNOISE, as shown in Figure 1, which first constructs a 3D noise space over the observable world region and distributes randomness throughout this space. Different views therefore become different observations of the same underlying noise world, rather than being generated from unrelated noise. This shared stochastic origin enables highly consistent observations across views.

However, constructing such a noise world is not straightforward. It requires distributing randomness over a 3D region under a finite budget while ensuring that the view-specific noise observed by each camera satisfies the native spatial noise prior expected by the video diffusion model. We first define the noise field over the observable world region jointly covered by the camera frusta along the trajectory. We allocate randomness according to a simple principle: nearby regions require finer detail, whereas distant regions require less detail. Following this principle, we populate the field with noise primitives of nonuniform density and spatial extent, assigning denser, finer support nearby and sparser, coarser support farther away. We then project the world-space randomness into view-specific 2D noise through a carefully designed whitening and covariance-compensation process. This process recovers spatially i.i.d. Gaussian noise within each view while retaining highly correlated noise across views that observe overlapping regions. The resulting noise encodes shared 3D structure through the existing diffusion noise interface without modifying the denoiser architecture. We also extend WORLDNOISE to long-horizon video generation by allowing the noise world to persist and grow as exploration proceeds. To maintain consistency with previously generated regions and address occlusions, we develop two geometric carving rules for newly introduced noise. These rules allow new support to respect the geometry of explored regions as the noise world expands. To validate the effectiveness of our method, we integrate it into two video diffusion backbones, Wan2.1 (Wan et al., 2025) and LTX-2.5 (Lightricks, 2026), demonstrating its model-agnostic applicability, and evaluate it across different levels of camera-motion complexity.

To summarize, the key contributions of this work are as follows: (i) We introduce WORLDNOISE, a framework that structures diffusion noise as a latent representation of the 3D world, enabling video diffusion models to observe a shared noise world from different camera viewpoints. (ii) We design a suite of algorithms for constructing the 3D noise field and projecting it into view-specific noise that preserves the native spatial prior and cross-view correlations. We further extend WORLDNOISE to long-horizon generation with a persistent, growing noise world. (iii) Extensive experiments across different diffusion backbones and under both standard and challenging camera motions demonstrate the effectiveness of WORLDNOISE.

2 Related Work

Camera-Controlled Video Generation and World Modeling. Camera-controlled video generation aims to synthesize coherent scene observations along specified trajectories. Camera conditioning follows three main strategies according to how geometry enters the generative process. One line of work relies on explicit visual conditioning, translating camera geometry into rendered or warped appearance cues. ViewCrafter (Yu et al., 2024) reconstructs scene point clouds and renders target views as RGB guidance, while GEN3C (Ren et al., 2025) maintains a 3D cache of observed content and conditions generation on its renderings. CamTrol (Hou & Chen, 2024) and Latent-Reframe (Zhou et al., 2025) further exploit geometric transformations of images or intermediate latents to steer the sampling process. These methods provide dense appearance guidance whose reliability depends on the reconstructed geometry and its projection into target views. Another line of work uses learned camera-feature conditioning, encoding camera parameters into features that interact with the denoiser. MotionCtrl (Wang et al., 2024) injects camera extrinsics into temporal layers, whereas CameraCtrl (He et al., 2025) uses a dedicated encoder to extract features from per-pixel Plücker rays. These representations allow the model to learn how camera motion governs changes in visual content. Geometry-aware attention provides a further route by incorporating multi-view relationships into token interactions through positional encoding or explicit geometric constraints. UCPE (Zhang et al., 2026) combines relative ray encoding with absolute orientation encoding, ReRoPE (Li et al., 2026) repurposes low-frequency temporal RoPE bands for relative camera information, and SCoPE (Yin et al., 2026) introduces Plücker sightlines into attention queries and keys while preserving pretrained RoPE. CamCo (Xu et al., 2024) and CamI2V (Zheng et al., 2024) impose epipolar constraints on cross-view attention to guide feature aggregation according to multi-view geometry. These designs place camera geometry directly in the computation of attention scores and the selection of cross-view interactions. Advances in camera control have further promoted the development of interactive world models, extending viewpoint-conditioned synthesis to sustained exploration of generated environments; we discuss these models in Appendix A.

Structured Noise for Diffusion Models. Previous work has explored correlated noise in diffusion models (Wu et al., 2024; Ge et al., 2023; Qiu et al., 2024) . These methods organize randomness across time, introducing temporal correlations in noise through operations such as low-frequency refinement, noise sharing across frames, and shuffled reuse of noise maps, to improve visual continuity across generated frames. Go-with-the-Flow (Burgert et al., 2025) further introduces spatial correspondences into the noise. It warps Gaussian noise according to optical flow to control motion while preserving spatial Gaussianity, enabling local object motion control and reference-based camera movement control. Recent studies further encode geometry in diffusion noise. CameraNoise (Zhao et al., 2026) derives geometry-guided reprojection flow from camera parameters and uses it to propagate noise between adjacent frames. It thereby encodes camera motion through the sequential transport of a 2D noise state. UniCaMo (Vu et al., 2026) transports reference-frame noise along sparse 3D point tracks and samples a panoramic Gaussian noise map on a predefined sphere to supply noise for newly visible regions, enabling joint control of object and camera motion. These geometry-guided noise methods use image-plane transport or panoramic noise on a predefined sphere, rather than a volumetric noise field throughout the observable world region.

3 Method

3.1 Stochastic Support for World Construction

Video diffusion models generate structured visual patterns by progressively transforming Gaussian noise into a coherent video (Ho et al., 2022). The initial noise is sampled independently across temporal and spatial dimensions (Figure 2(a)). In the flow-matching formulation, generation evolves from this random state by integrating a learned velocity field (Lipman et al., 2022):

$$\epsilon_{\ell,h,w} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad \mathbf{x}_{\tau=1} = \boldsymbol{\epsilon}, \qquad d\mathbf{x}_\tau/d\tau = v_\theta(\mathbf{x}_\tau, \tau), \quad \tau : 1 \to 0. \tag{1}$$

where ℓ, h, w index latent time and spatial locations. While conventional diffusion leaves the random source itself unstructured, prior work has shown that introducing temporal or spatial correlations into the initial noise can substantially influence temporal coherence and motion controllability (Burgert et al., 2025; Zhao et al., 2026; Vu et al., 2026). This raises a more central question for world modeling: what kind of stochastic support is suitable for constructing a persistent world? One simple idea is to lift 2D noise to a prescribed pseudo-depth (Figure 2(b)) and establish geometric correspondence

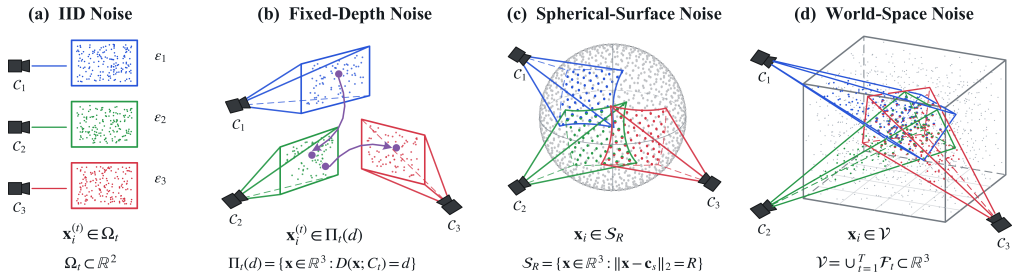

Figure 2: **Geometric supports of stochastic primitives.** Different noise constructions define stochastic primitives over different geometric supports, from independent 2D view spaces to fixed-depth planes, a shared spherical surface, and finally a shared 3D world volume.

through reprojection. Here, correlations are only established by sequentially propagating noise between adjacent frames. To solve this limitation, a shared spherical-surface field instead allows different viewpoints to access the same stochastic identities (Figure 2(c)). The sphere provides a global reference across views, but each viewing ray is restricted to a single surface depth.

These observations suggest that noise for world construction should satisfy two properties: (i) sharing across all views, and (ii) distribution across multiple depths in the world. Therefore, we argue that noise for world construction should be defined directly in the observable 3D world space (Figure 2(d)). Let $\mathcal{Z}$ denote a stochastic field over $\mathcal{V} \subset \mathbb{R}^3$, shared by all viewpoints, and let $\mathcal{F}_t$ denote the viewing frustum of camera C_t. Each view obtains its noise $\boldsymbol{\epsilon}_t$ by observing the shared field through the corresponding frustum:

$$\mathcal{Z} : \mathcal{V} \to \mathbb{R}^C, \quad \boldsymbol{\epsilon}_t = \mathcal{R}(\mathcal{Z}; \mathcal{F}_t), \quad \mathrm{Cov}(\boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_s) \begin{cases} = 0, & \mathcal{F}_t \cap \mathcal{F}_s = \varnothing, \\ \neq 0, & \mathcal{F}_t \cap \mathcal{F}_s \neq \varnothing. \end{cases} \tag{2}$$

Thus, non-overlapping views remain independent, while overlapping views become correlated through their shared world-space randomness.

3.2 World-Space Noise Construction

In this section, we detail how to construct noise directly in world space, which forms the core of our WorldNoise framework. Constructing such a noise world is not straightforward: it must represent stochastic structure over a large 3D space under a finite budget, while its view-specific projections must remain compatible with the native noise prior of the pretrained video diffusion model.

World-space stochastic support. Given a camera trajectory $\{C_t\}_{t=1}^{T}$, each camera C_t defines a finite viewing frustum $\mathcal{F}_t$. For a world point $\mathbf{x}$, let $\Pi(\mathbf{x}; C_t)$ and $D(\mathbf{x}; C_t)$ denote its image projection and camera-space depth under C_t, respectively. Let Ω_h be the $H_h \times W_h$ raster domain, with $0 < z_{\min} < z_{\max}$. The observable world introduced in Section 3.1 is therefore:

$$\mathcal{F}_t = \left\{ \mathbf{x} \in \mathbb{R}^3 : \Pi(\mathbf{x}; C_t) \in \Omega_h,\ z_{\min} \le D(\mathbf{x}; C_t) \le z_{\max} \right\}, \quad \mathcal{V} = \bigcup_{t=1}^{T} \mathcal{F}_t. \tag{3}$$

The remaining question is how to discretize $\mathcal{V}$ into a finite set of world-space noise primitives. As shown in Figure 3(a), we adopt a non-uniform discretization motivated by the way visual information is perceived across depth: nearby regions typically convey richer spatial detail, while distant regions provide progressively coarser information. Specifically, we sample uniformly in image coordinates and logarithmic depth. For a scrambled Sobol sample (Sobol', 1967) $\mathbf{a} = (a_0, a_1, a_2) \in [0, 1)^3$,

$$u = a_0 W_h, \qquad v = a_1 H_h, \qquad z = z_{\min} \exp\!\left(a_2 \log\left(\frac{z_{\max}}{z_{\min}}\right)\right), \tag{4}$$

and the sample at (u, v, z) is then unprojected to a primitive position $\mathbf{x}_i$ in the shared world space. We next determine the spatial support of each primitive from its local world-space density. For camera C_t, the density at a visible primitive location $\mathbf{x}_i$ is

$$\rho_t(\mathbf{x}_i) = \frac{N_t |f_{x,t} f_{y,t}|}{H_h W_h \log(z_{\max}/z_{\min})\, D(\mathbf{x}_i; C_t)^3}, \qquad \rho(\mathbf{x}_i) = \sum_{t : \mathbf{x}_i \in \mathcal{F}_t} \rho_t(\mathbf{x}_i), \tag{5}$$

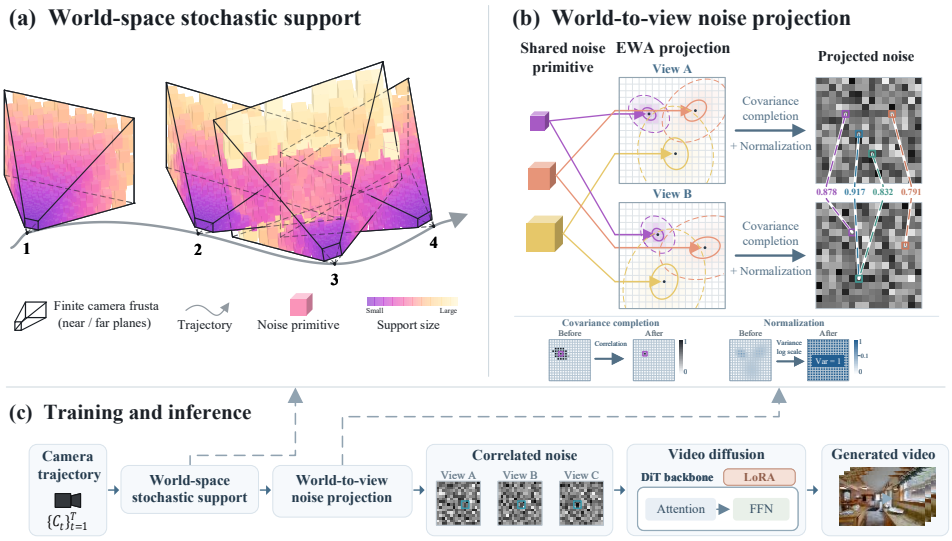

Figure 3: **WORLDNOISE construction and integration.** (a) World-space stochastic support. (b) World-to-view noise projection with covariance completion and normalization. (c) WORLDNOISE replaces the native noise input for video diffusion training and inference.

where N_t denotes the sample count for camera C_t, and $f_{x,t}, f_{y,t}$ are its focal lengths in raster pixels. We assign each primitive a spatial extent h_i and an independent, persistent Gaussian identity $\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_C)$, forming the world-space noise field $\mathcal{P}$:

$$h_i = \min\left(\rho(\mathbf{x}_i)^{-1/3}, h_i^{\text{proj}}\right), \qquad \mathcal{P} = \{(\mathbf{x}_i, h_i, \mathbf{z}_i)\}_{i=1}^{N}. \tag{6}$$

Here, $\rho(\mathbf{x}_i)^{-1/3}$ follows the local inter-primitive spacing, and h_i^{proj} prevents the primitive from projecting to an excessively large region in any observing view. The Details are provided in Appendix B.

World-to-view noise projection. Given the world-space primitives $\mathcal{P}$, we project each primitive into a view-specific footprint using elliptical weighted average (EWA) splatting (Zwicker et al., 2002), shown in Figure 3(b). Our goal is to preserve the shared world-space stochastic identities while recovering the spatial noise prior expected by the video diffusion model. For primitive i under camera C_t, its 3D spatial extent is projected to the image plane with covariance

$$\mathbf{\Sigma}_{ti} = J_{ti}(h_i^2 I_3/12)J_{ti}^{\top}, \tag{7}$$

where J_{ti} is the perspective-projection Jacobian at $\mathbf{x}_i$, and $h_i^2 I_3/12$ describes the spatial extent of the primitive (details are provided in Appendix B). The EWA footprint defines weights w_{tij}, indicating the contribution of primitive i to pixel j in view t. Directly splatting the shared feature $\mathbf{z}_i$ induces within-view covariance $\mathbf{w}_{ti}\mathbf{w}_{ti}^{\top}$, so we add an independent residual:

$$\mathbf{g}_{ti} = \mathbf{w}_{ti}\mathbf{z}_i^{\top} + \mathbf{r}_{ti}, \qquad \mathbf{r}_{ti}^{(c)} \sim \mathcal{N}\left(\mathbf{0}, \operatorname{diag}(\mathbf{w}_{ti}) - \mathbf{w}_{ti}\mathbf{w}_{ti}^{\top}\right). \tag{8}$$

We then accumulate all projected contributions and normalize their variance:

$$A_{tj} = \sum_i g_{tij}, \qquad E_{tj} = \sum_i w_{tij}, \qquad N_{tj} = \begin{cases} A_{tj}/\sqrt{E_{tj}}, & E_{tj} > 0, \\ \epsilon_{tj}, & E_{tj} = 0, \end{cases} \tag{9}$$

where $\epsilon_{tj} \sim \mathcal{N}(0, I_C)$. The resulting noise is spatially i.i.d. within each view while remaining correlated across views that share world primitives.

3.3 WORLDNOISE FOR VIDEO DIFFUSION AND LONG-HORIZON EXTENSION

WORLDNOISE for Camera-Controlled Video Generation. As shown in Figure 3(c), WORLDNOISE can be directly integrated into a pretrained video diffusion model by replacing the original i.i.d. Gaussian noise with the correlated noise. The denoising architecture and training objective remain

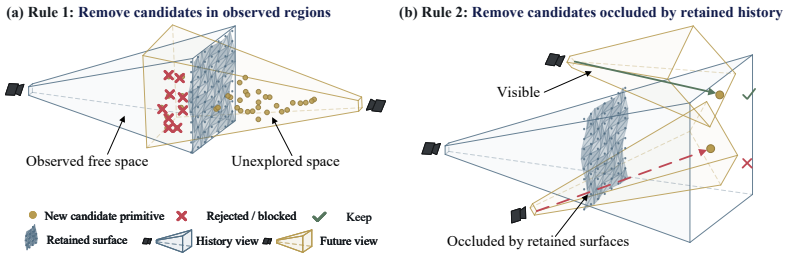

Figure 4: Two rules for **Geometric carving for long-horizon world expansion.**

unchanged, and we adapt the model using LoRA (Hu et al., 2022) to align it with the structured noise distribution.

Extension to long-horizon generation. For long-horizon generation, we maintain the previously observed world and introduce new noise primitives only in regions covered by the upcoming trajectory. At chunk k, retained history H_k is combined with a temporary volume V_k sampled from future camera frusta. As illustrated in Figure 4, two geometric carving rules filter the newly sampled primitives according to previously observed regions and visibility against retained surfaces. Across chunks, spatial memory retrieves four historical observations that best cover the upcoming views, while temporal memory retains the most recent 16 RGB frames. Future noise is rendered from $H_k \cup V_k$. After generation, reliable observations are committed to the world history and V_k is discarded. Further details are provided in Appendix C.

4 Experiments

4.1 Experimental Setting

Implementation Details. To evaluate the effectiveness and generality of the proposed WorldNoise, we integrate it into two video diffusion backbones, Wan2.1-14B (Wan et al., 2025) and LTX-2.5-22B (Lightricks, 2026). Both models are adapted using rank-32 LoRA with a learning rate of 1×10^{-5} for 10K training iterations on 8 NVIDIA A800 GPUs. During inference, we use 50 denoising steps for Wan2.1 and 30 denoising steps for LTX-2.5.

Datasets. We train our models on DL3DV-10K (Ling et al., 2024), which contains 10K long videos covering diverse real-world scenes and camera trajectories. For each training video, we use ViPE (Huang et al., 2025) to estimate the camera trajectory and per-frame depth maps, and Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) to generate the corresponding video caption. For evaluation, we construct two test sets with different levels of camera-motion complexity. We first sample 100 videos from the RealEstate10K test set (Zhou et al., 2018), a widely adopted benchmark for camera-controllable video generation. To further evaluate performance under more challenging camera trajectories, we additionally sample 100 videos from DL3DV-10K, which contains substantial camera motions. Short-video evaluations are performed on 81-frame clips at a spatial resolution of 832×480. Long-horizon evaluation uses DL3DV-10K sequences longer than 300 frames and is reported separately from the 81-frame comparisons.

Evaluation Metrics. We evaluate generated videos from three complementary aspects: camera controllability, geometric consistency, and perceptual video quality. We recover camera trajectories using GLOMAP (Pan et al., 2024) and evaluate camera controllability with rotation error (RotErr), translation error (TransErr), and relative pose errors in translation and rotation (RPE-t and RPE-r), following camera-trajectory evaluation protocols used in prior camera-controlled video generation (Zheng et al., 2024). We additionally report pose registration coverage (Reg.) to measure the fraction of frames for which camera poses can be successfully recovered. For geometric consistency, we use TSED (Yu et al., 2023), which measures epipolar consistency under the target camera geometry. For perceptual video quality, we report Motion Smoothness (MS), Dynamic Degree (DD), Aesthetic Quality (AQ), and Imaging Quality (IQ) from VBench (Huang et al., 2024). Detailed definitions and evaluation protocols are provided in Appendix D.

Baselines. We compare WORLDNOISE with representative state-of-the-art camera-controlled video generation methods, including SCoPE (Yin et al., 2026), GEN3C (Ren et al., 2025), and UCM (Xu et al., 2026), which introduce camera control through positional encoding, explicit geometric conditioning, and time-aware positional encoding warping, respectively. To more directly evaluate the effect of noise construction, we additionally instantiate the noise construction strategies introduced in CameraNoise (Zhao et al., 2026) and UniCaMo (Vu et al., 2026) for camera-controlled video generation. These variants are trained on the same dataset as WORLDNOISE and evaluated under the same protocol. Finally, we compare with recent world-modeling systems, Lyra 2.0 (Shen et al., 2026) and SANA-WM (Zhu et al., 2026), to evaluate the long-horizon extension of WORLDNOISE against dedicated world-generation systems.

4.2 COMPARISON WITH CAMERA-CONTROLLED VIDEO GENERATION METHODS

We directly compare WORLDNOISE with state-of-the-art camera-controlled video generation methods. We integrate WORLDNOISE into Wan2.1 and LTX-2.5 and evaluate them on RealEstate10K and DL3DV, representing standard and challenging camera-motion settings, respectively. As summarized in Table 1, incorporating WORLDNOISE substantially strengthens the camera controllability of the underlying video diffusion backbones, matching or surpassing specialized state-of-the-art camera-

Table 1: Comparison with camera-controlled video generation methods.

	Method	Camera Control				Geometry		VBench			
		RotErr↓	TransErr↓	RPE-t↓	RPE-r↓	TSED↑	Reg.↑	MS↑	DD↑	AQ↑	IQ↑
RealEstate10K	SCoPE	2.324	0.1780	0.0195	0.136	0.9510	**1.0000**	0.9903	0.6900	**0.5385**	0.7361
	Gen3C	2.099	0.2716	**0.0152**	0.147	0.9644	0.9796	0.9904	0.5500	0.5118	0.7042
	UCM	2.006	**0.1199**	0.0180	0.164	**0.9865**	0.9995	0.9918	0.5900	0.5168	0.7305
	Wan	9.016	0.5372	0.0413	0.489	0.8010	0.9733	0.9810	0.7400	0.4927	0.7127
	Wan + Ours	**1.940**	0.2422	0.0216	**0.128**	0.9779	0.9996	0.9907	0.6800	0.5242	**0.7448**
	LTX	4.860	0.2880	0.0225	0.230	0.9323	**1.0000**	0.9939	0.6400	0.5151	0.7143
	LTX + Ours	2.084	0.2255	0.0157	0.145	0.9732	**1.0000**	**0.9971**	**0.8300**	0.5322	0.7093
DL3DV	SCoPE	8.552	**0.0960**	0.0996	1.374	0.3613	0.9875	0.9702	**1.0000**	**0.4877**	0.6507
	Gen3C	15.092	0.1676	0.0830	1.566	0.5000	0.8245	0.9645	**1.0000**	0.4342	0.5080
	UCM	19.965	0.2131	0.0860	1.698	**0.5607**	0.9565	0.9710	**1.0000**	0.4578	0.6088
	Wan	43.815	0.4672	0.1606	2.661	0.2586	0.9363	0.9555	0.9219	0.4713	0.6613
	Wan + Ours	7.664	0.1806	**0.0778**	**1.033**	0.4865	0.9666	0.9636	**1.0000**	0.4805	**0.6770**
	LTX	37.649	0.3955	0.1169	2.235	0.3371	0.9898	0.9866	0.9063	0.4801	0.6519
	LTX + Ours	**7.062**	0.1514	0.1024	1.409	0.5355	**0.9947**	**0.9887**	**1.0000**	0.4806	0.6540

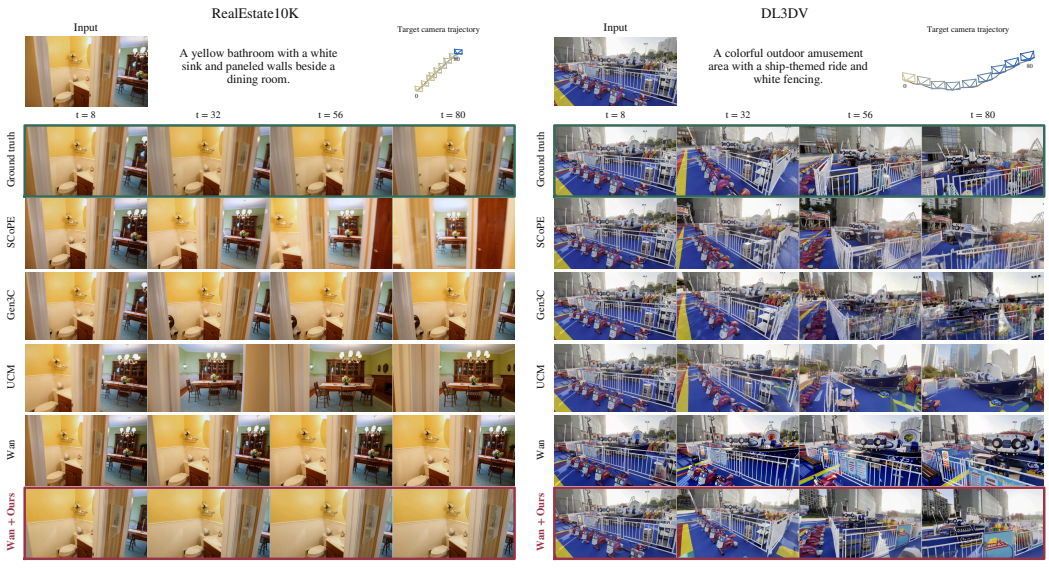

Figure 5: Qualitative comparison on RealEstate10K and DL3DV.

Table 2: Comparison of noise construction strategies.

	Method	Camera Control				Geometry		VBench			
		RotErr↓	TransErr↓	RPE-t↓	RPE-r↓	TSED↑	Reg.↑	MS↑	DD↑	AQ↑	IQ↑
RealEstate10K	CameraNoise	7.249	**0.2381**	0.0247	0.347	0.9045	0.9973	0.9903	**0.6800**	0.5173	0.7349
	UniCaMo	7.316	0.4825	0.0300	0.388	0.9286	0.9953	0.9904	0.5500	0.5204	0.7330
	WORLDNOISE	**1.940**	0.2422	**0.0216**	**0.128**	**0.9779**	**0.9996**	**0.9907**	**0.6800**	**0.5242**	**0.7448**
DL3DV	CameraNoise	36.353	0.3426	0.1465	2.706	0.1648	0.9279	0.9510	0.9844	0.4784	0.6756
	UniCaMo	37.701	0.3712	0.1346	2.486	0.2715	0.9192	0.9635	0.9531	**0.4926**	**0.6962**
	WORLDNOISE	**7.664**	**0.1806**	**0.0778**	**1.033**	**0.4865**	**0.9666**	**0.9636**	**1.0000**	0.4805	0.6770

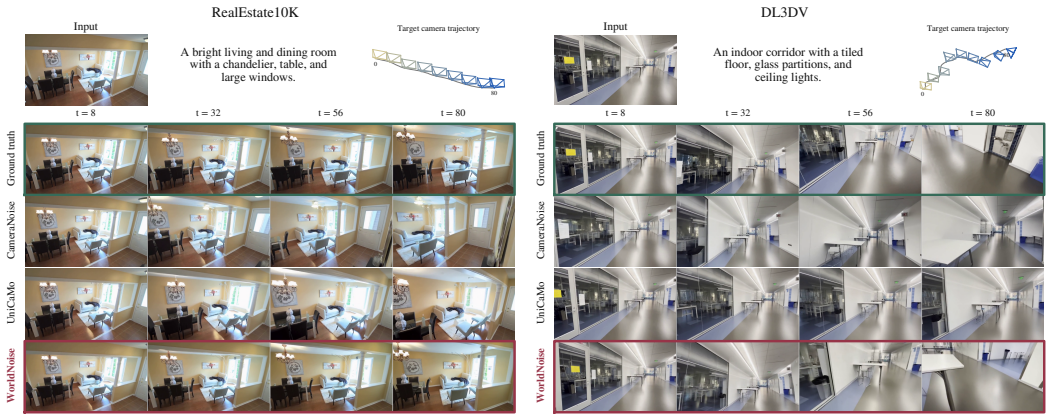

Figure 6: Qualitative comparison of noise construction strategies.

control methods under both standard and challenging camera motions. Qualitative comparisons are shown in Figure 5. WORLDNOISE achieves accurate camera control while maintaining strong visual quality, which is consistent with its design: geometry is introduced through the noise space while preserving the native diffusion architecture.

4.3 COMPARISON OF NOISE CONSTRUCTION STRATEGIES

To more directly isolate the effect of noise construction, we adapt the strategies introduced in CameraNoise (Zhao et al., 2026) and UniCaMo (Vu et al., 2026) to the camera-control setting. For a fair comparison, we retrain all variants on Wan2.1 using the same dataset and evaluate them under the same protocol. As shown in Table 2, WORLDNOISE achieves the strongest overall camera-control performance on both RealEstate10K and DL3DV, with particularly pronounced gains on camera and geometric metrics under the more challenging motions in DL3DV. These results further support the effectiveness of grounding stochastic correspondence directly in world space. Qualitative results in Figure 6 further reveal this difference. Under challenging camera motions, CameraNoise and UniCaMo show less faithful spatial evolution, consistent with their fixed-depth or spherical-surface noise representations that cannot fully capture multi-depth parallax. In contrast, WORLDNOISE maintains more accurate camera motion and spatial relationships across views.

4.4 EXTENSION TO LONG-HORIZON VIDEO GENERATION

We further extend WORLDNOISE on Wan2.1 to support long-horizon video generation. By maintaining and expanding the shared world-space noise as exploration proceeds, the extended model can generate videos beyond the short-window setting while preserving camera controllability and scene consistency. We compare against dedicated world-generation systems, including SANA-WM (Zhu et al., 2026) and Lyra 2.0 (Shen et al., 2026) in Table 3, with qualitative comparisons in Figure 7. The results show that, despite its simple extension from a standard video diffusion backbone, WORLDNOISE achieves comparable long-horizon generation performance to these carefully designed world models.

Table 3: Long-horizon video generation on DL3DV.

Method	Camera Control				Geometry		VBench			
	RotErr↓	TransErr↓	RPE-t↓	RPE-r↓	TSED↑	Reg.↑	MS↑	DD↑	AQ↑	IQ↑
SANA-WM	15.581	**0.1871**	0.0986	1.434	0.3216	0.6025	**0.9657**	**1.0000**	0.4729	**0.6372**
Lyra 2.0	17.664	0.2817	0.1135	1.543	0.3994	0.7105	0.9648	**1.0000**	**0.4767**	0.6189
WORLDNOISE	**14.570**	0.1915	**0.0818**	**1.363**	**0.4026**	**0.7674**	0.9637	**1.0000**	0.4709	0.6337

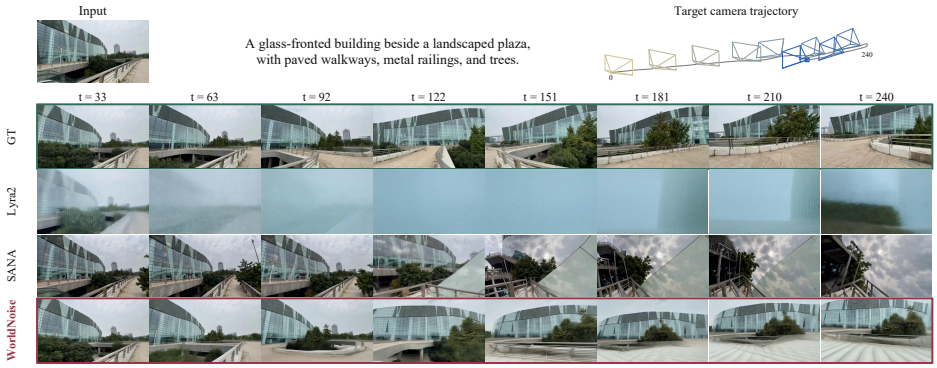

Figure 7: Qualitative comparison of long-horizon video generation.

4.5 ABLATION STUDIES

We ablate two key design choices in the construction of WORLDNOISE: the point sampling budget and the depth sampling strategy. We retrain all variants for about 5000 steps and test on DL3DV. As shown in Table 4, we evaluate three sampling budgets under both uniform-depth and log-depth sampling. **Point sampling budget.** Increasing the number of noise primitives consistently improves the camera-control metrics. A denser set of primitives provides finer spatial resolution in the stochastic world, allowing cross-view relationships to be represented more precisely and thereby improving camera controllability. **Log-depth sampling.** Under the same sampling budget, log-depth sampling consistently outperforms uniform-depth sampling. By allocating more primitives to nearby regions and fewer to distant regions, log-depth sampling concentrates the stochastic correlations on spatial regions that require finer geometric resolution, leading to more effective world-space noise construction.

Table 4: Ablation of sampling budget and log-depth sampling.

Points	Log-depth	Camera Control				Geometry		VBench			
		RotErr↓	TransErr↓	RPE-t↓	RPE-r↓	TSED↑	Reg.↑	MS↑	DD↑	AQ↑	IQ↑
1M	w/o	11.295	0.2318	0.1666	1.449	0.4554	**0.9641**	**0.9600**	**1.0000**	0.4838	**0.6943**
	w/	**11.041**	**0.1958**	**0.1608**	**1.379**	**0.4570**	**0.9641**	0.9586	**1.0000**	**0.4852**	0.6851
4M	w/o	11.357	0.2288	0.1855	1.311	0.4755	0.9745	0.9605	**1.0000**	0.4801	0.6900
	w/	**10.264**	**0.1913**	**0.1583**	**1.257**	**0.4849**	**0.9749**	**0.9608**	**1.0000**	**0.4912**	**0.6968**
8M	w/o	8.833	0.2108	0.1631	**1.159**	0.4743	0.9506	**0.9617**	**1.0000**	0.4853	0.6982
	w/	**8.555**	**0.1884**	**0.1571**	1.217	**0.4779**	**0.9684**	0.9603	**1.0000**	**0.4859**	**0.7040**

5 CONCLUSION AND LIMITATIONS

We introduce WORLDNOISE, which constructs a shared stochastic world before video synthesis, with persistent multi-depth primitives establishing spatial identity and Gaussian-preserving projection translating that identity into cross-view noise correlations. The resulting noise can be used through the existing diffusion noise interface without modifying the denoiser architecture, and can be extended to long-horizon generation by allowing the noise world to persist and grow. Experiments on Wan2.1 and LTX-2.5 demonstrate its effectiveness under both standard and challenging camera motions. These results reveal diffusion noise as a latent carrier of 3D structure for 3D-aware generative world models.

AI USE STATEMENT

We use AI-based tools to assist with language polishing and improve the readability of the manuscript. All technical ideas, methodology design, experimental settings, analysis, and conclusions are developed, verified, and finalized by the authors.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Ryan Burgert, Yuancheng Xu, Wenqi Xian, Oliver Pilarski, Pascal Clausen, Mingming He, Li Ma, Yitong Deng, Lingxiao Li, Mohsen Mousavi, Michael Ryoo, Paul Debevec, and Ning Yu. Go-with-the-flow: Motion-controllable video diffusion models using real-time warped noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13–23, 2025. doi: 10.1109/CVPR52734.2025.00011. URL https://openaccess.thecvf.com/content/CVPR2025/html/Burgert_Go-with-the-Flow_Motion-Controllable_Video_Diffusion_Models_Using_Real-Time_Warped_Noise_CVPR_2025_paper.html.

Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve your own correlation: A noise prior for video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22930–22941, 2023. doi: 10.1109/ICCV51070.2023.02096. URL https://openaccess.thecvf.com/content/ICCV2023/html/Ge_Preserve_Your_Own_Correlation_A_Noise_Prior_for_Video_Diffusion_ICCV_2023_paper.html.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3D geometric perception. *arXiv preprint arXiv:2508.10934*, 2025. URL https://arxiv.org/abs/2508.10934.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21807–21818. IEEE, 2024.

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*, 2026. URL https://arxiv.org/abs/2602.08068.

Lightricks. LTX-2.5: LTX's latest AI open-source foundation model. Official model page, 2026. URL https://ltx.io/model/ltx-2-5.

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

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22160–22169. IEEE, 2024.

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

Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, pp. 58–77. Springer, 2024.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. Google DeepMind, 2024. URL https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/.

Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/15ce8e7afe5ee95bad56e3b9be28d3d1-Abstract-Conference.html.

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. doi: 10.1109/CVPR52734.2025.00574. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ren_GEN3C_3D-Informed_World-Consistent_Video_Generation_with_Precise_Camera_Control_CVPR_2025_paper.html.

Tianchang Shen, Sherwin Bahmani, Kai He, Sangeetha Grama Srinivasan, Tianshi Cao, Jiawei Ren, Ruilong Li, Zian Wang, Nicholas Sharp, Zan Gojcic, Sanja Fidler, Jiahui Huang, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra 2.0: Explorable generative 3d worlds. *arXiv preprint arXiv:2604.13036*, 2026. URL https://arxiv.org/abs/2604.13036.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *U.S.S.R. Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9. URL https://www.mathnet.ru/eng/zvmmf7334.

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

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

Long Vu, Tan Ngo, Animesh Karnewar, Amir Habibian, Binh-Son Hua, Hung Bui, Minh Hoai Nguyen, and Phong Nguyen-Ha. Track the noise, move the world: 3d-grounded motion-consistent noise for controllable video generation. *arXiv preprint arXiv:2607.02798*, 2026. URL https://arxiv.org/abs/2607.02798.

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

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pp. 378–394, 2024. doi: 10.1007/978-3-031-72646-0_22. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/423_ECCV_2024_paper.php.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. CamCo: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. URL https://arxiv.org/abs/2406.02509.

Tian-Xing Xu, Zi-Xuan Wang, Guangyuan Wang, Li Hu, Zhongyi Zhang, Peng Zhang, Bang Zhang, and Song-Hai Zhang. Ucm: Unified modeling of camera control and memory with time-aware positional encoding warping for world models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Shan Ying, and Kai Han. Scope: Sightline-coordinate positional encoding for video diffusion transformers, 2026. URL https://arxiv.org/abs/2606.27345.

Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2304.10700.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. URL https://arxiv.org/abs/2409.02048.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.07237.

Haoyu Zhao, Jiaxi Gu, Haoran Chen, Qingping Zheng, Yeying Jin, Hongyi Yang, Junqi Cheng, Yuang Zhang, Zenghui Lu, Huan Yu, Jie Jiang, Peng Shu, Zuxuan Wu, and Yu-Gang Jiang. Cameranoise: Enabling faithful camera control in video diffusion through geometry-flow-guided noise warping. In *Proceedings of the Forty-Third International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2605.30774.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. CamI2V: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024. URL https://arxiv.org/abs/2410.15957.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018. doi: 10.1145/3197517.3201323. URL https://google.github.io/realestate10k/.

Zhenghong Zhou, Jie An, and Jiebo Luo. Latent-reframe: Enabling camera control for video diffusion models without training. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12779–12789. IEEE, 2025.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. SANA-WM: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026. URL https://arxiv.org/abs/2605.15178.

Matthias Zwicker, Hanspeter Pfister, Jeroen van Baar, and Markus Gross. EWA splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3):223–238, 2002. URL https://www.merl.com/publications/TR2002-49.

A Interactive World Models

SANA-WM (Zhu et al., 2026) combines hybrid linear attention with dual-branch camera control to support minute-scale generation. HY-WorldPlay (Sun et al., 2025) integrates discrete actions and continuous camera poses with reconstituted context memory for long-term geometric consistency. LingBot-World (Team et al., 2026) introduces action conditioning through adaptive normalization and adapts video diffusion to interactive generation through causal attention and few-step distillation. Lyra 2.0 (Shen et al., 2026) maintains per-frame geometry to retrieve relevant history and establish cross-view correspondences, and trains with self-augmented histories to reduce temporal drift. Although these approaches improve camera controllability and scene consistency, they still rely on explicit visual guidance, additional learnable components, or specialized encoding and architectural designs to represent camera and world structure.

B World-Space Noise: Construction and Statistical Properties

Shared support and sampling. All cameras use one world frame and a common depth scale. Samples from their finite frusta are unprojected into $\mathcal{V}$ and assigned persistent Gaussian identities $\mathbf{z}_i$. The Sobol construction in Equation 4 allocates samples uniformly in image coordinates and logarithmic depth. Its depth density is proportional to z^{-1}, while the unprojection volume Jacobian is $z^2/|f_{x,t}f_{y,t}|$. Their ratio gives the z^{-3} world-space density in Equation 5. Thus the nominal spacing grows linearly with depth, matching the coarser world-space resolution of distant image regions. Densities from overlapping frusta add. When first-view depth is available, surface anchors supplement the volume, and candidates in observed free space or close to those surfaces are removed.

Primitive extent and projection. The extent h_i combines local spacing with a projected-size bound. With J_{ti} the perspective Jacobian, the implementation uses

$$h_i^{\mathrm{proj}} = \frac{\eta}{\sqrt{3\lambda_i^*}}, \qquad \lambda_i^* = \max_{t:\mathbf{x}_i\in\mathcal{F}_t} \lambda_{\max}(J_{ti}J_{ti}^\top), \qquad 0<\eta<1, \tag{10}$$

where η is a safety margin. This bounds the footprint in observing views. EWA rasterization (Zwicker et al., 2002) adds the pixel-box covariance to the geometric covariance in Equation 7: $\widetilde{\mathbf{\Sigma}}_{ti} = \mathbf{\Sigma}_{ti} + I_2/12$. Discrete Gaussian weights are normalized over the complete footprint before off-image targets are discarded. Consequently, $w_{tij} \geq 0$ and $q_{ti} = \sum_j w_{tij} \leq 1$.

Covariance completion. Let $\Lambda = \mathrm{diag}(\mathbf{w}_{ti})$, $\mathbf{v} = \sqrt{\mathbf{w}_{ti}}$, and $\beta = (1+\sqrt{1-q_{ti}})^{-1}$. For a matrix $\boldsymbol{\xi}_{ti}$ of independent standard Gaussian entries, independent of primitive identities, the residual is

$$\mathbf{r}_{ti} = \Lambda^{1/2}(I-\beta\mathbf{v}\mathbf{v}^\top)\boldsymbol{\xi}_{ti}, \qquad \mathrm{Cov}(\mathbf{g}_{ti}^{(c)}) = \mathrm{diag}(\mathbf{w}_{ti}). \tag{11}$$

The identity $1-2\beta+q_{ti}\beta^2=0$ yields the residual covariance in Equation 8. Hence primitive i contributes variance w_{tij} to pixel j, giving the normalization $\sqrt{E_{tj}} = \sqrt{\sum_i w_{tij}}$.

Proposition 1 (Within-view Gaussianity). Condition on geometry $\mathcal{G}$, including cameras, support, and visibility. Independent Gaussian identities, residuals, and empty-pixel fills make each view spatially i.i.d. standard Gaussian. Indeed, each primitive has diagonal spatial covariance; independent accumulation gives variance E_{tj}, and normalization gives unit variance. Joint Gaussianity then implies pixel independence.

Proposition 2 (Cross-view covariance). For occupied pixels in separately rendered views with independent residuals,

$$\mathrm{Cov}(N_{tj}, N_{sk} \mid \mathcal{G}) = \frac{\sum_i w_{tij}w_{sik}}{\sqrt{E_{tj}E_{sk}}}\, I_C. \tag{12}$$

Only shared $\mathbf{z}_i$ terms survive the cross-view covariance, proving the expression. Correlation therefore follows shared visible primitives; frustum overlap identifies where sharing is possible. Independent empty fills have zero cross-view covariance. Consecutive identical cameras reuse the complete noise frame, a separate case from independent-residual rendering.

Algorithm 1 WORLDNOISE: world construction and projection

Require: cameras $\{C_t\}$, finite depth bounds, point budget, optional first-view depth
Ensure: projected noise $\{N_{tj}\}$
express cameras and depth in a shared world frame and scale
sample the frustum union by Eq. 4; add optional surface anchors
remove candidates in observed free space and near anchored surfaces
set h_i by Eq. 6; draw persistent $\mathbf{z}_i \sim \mathcal{N}(0, I_C)$
for all views t **do**
    project primitives and normalize EWA weights before image clipping
    construct compensated contributions by Eq. 11
    accumulate A_{tj} and E_{tj}; normalize and fill empty pixels by Eq. 9
end for
reuse complete projections for consecutive identical cameras
return $\{N_{tj}\}$

C VIDEO DIFFUSION AND LONG-HORIZON GENERATION

Noise interface and model adaptation. Projected noise is reduced to the backbone's latent lattice by variance-preserving pooling. For a non-overlapping block $\mathcal{B}$, let d_{tj} be the relative accumulated weight of occupied pixel j. Then

$$L_{t\mathcal{B}} = \frac{\sum_{j\in\mathcal{B}} d_{tj} N_{tj}}{\sqrt{\sum_{j\in\mathcal{B}} d_{tj}^2}}. \tag{13}$$

Proposition 1 makes the inputs independent with unit variance, so the squared-coefficient normalization preserves that variance. The long-video renderer excludes empty pixels from pooling and independently fills wholly empty blocks; the single-window initializer retains its independent empty-pixel fills. Stacking the latent cells produces $\boldsymbol{\epsilon}$, which replaces native noise in Wan and LTX. LoRA (Hu et al., 2022) adapts the pretrained model to the new cross-frame distribution while retaining its denoising architecture and objective. The flow-matching path remains $\mathbf{x}_\tau = (1-\tau)\mathbf{x}_0 + \tau\boldsymbol{\epsilon}$, with target velocity $\boldsymbol{\epsilon} - \mathbf{x}_0$. Each endpoint is constructed once and held fixed during denoising.

Retained history and temporary volume. At chunk k, H_k stores retained surface primitives and their Gaussian identities. All observations share the first-camera coordinate frame and a fixed world scale. Reliable observations can update the finite depth interval used to sample new regions. Temporary V_k is sampled from the upcoming camera frusta using the log-depth construction of Appendix B. History surfaces use normal-oriented footprints; temporary volume primitives retain isotropic extents. Both use the same covariance completion and normalization. Retained observations keep their identities, while temporary primitives receive separate identities shared within the current chunk.

Geometric carving. The two rules in Figure 4 remove redundant or occluded temporary support. First, candidates in observed free space or close to retained surfaces are discarded. Second, the open camera-to-candidate segment is tested against trusted historical surfaces, producing a per-view visibility mask. A candidate survives if it remains visible in at least one future view. Historical geometry is built from valid depth regions, excluding discontinuities. These tests apply to V_k; retained H_k continues to be rendered with its own visibility and surface-orientation checks. Future target depth is unavailable during this construction.

Spatial and temporal context. Each chunk combines an initial anchor, four retrieved historical images, and the most recent sixteen RGB frames. Spatial retrieval greedily selects images that add coverage of the upcoming views; temporal memory preserves recent motion and appearance. The Wan implementation encodes these as $A_1/S_4/T_4$, followed by twelve future latent slots F_{12} generating 48 RGB frames. Retrieved images are encoded independently, and temporal frames are encoded jointly. Retrieval chooses visual context; the noise renderer uses the complete retained history. Writing $\mathcal{C}_M^{(k)}$ and $\mathcal{C}_F^{(k)}$ for memory and future cameras, and including latent reduction in $\mathcal{R}$, matched noise is

$$\boldsymbol{\epsilon}_M^{(k)} = \mathcal{R}(H_k; \mathcal{C}_M^{(k)}), \qquad \boldsymbol{\epsilon}_F^{(k)} = \mathcal{R}(H_k \cup V_k; \mathcal{C}_F^{(k)}). \tag{14}$$

All slots share a diffusion time. Memory slots are reconstructed from their known clean latents and matched noise at each step; only future slots receive loss and scheduler updates. Attention lets future slots access all context while preventing memory from reading the unknown future. The endpoint and history snapshot stay fixed until the chunk is completed.

Observation validation and update. After generation, completed RGB frames provide temporal context for the next chunk. Depth Anything 3 (Lin et al., 2025) estimates geometry from selected generated keyframes. Confidence and consistency with the first-frame and recent reliable references determine which observations can enter the world. Accepted surfaces update the history,

$$H_{k+1} = \mathrm{Retain}(H_k \cup \Delta H_k), \tag{15}$$

while V_k is discarded. New observations receive new identities; previously retained primitives preserve theirs. The same accepted observations update free-space evidence and retrieval records. Failed depth validation preserves generated RGB without changing geometry. History storage is bounded and preserves the initial anchor. Thus the stochastic support grows with exploration within the retention policy, and adding support can change projected mixtures even when existing identities stay fixed.

Training with generated history. Training progresses from initial chunks to completed ground-truth histories and then self-augmented histories. Self-augmentation makes one velocity prediction for the preceding chunk, decodes its estimated clean frames, and commits reliable geometry before constructing the current chunk. Recent generated frames also supply temporal context. This history-generation branch runs without gradients; the current future prediction receives the supervised update. The memory layout and future-only objective remain unchanged across stages.

Algorithm 2 WORLDNOISE: generation with retained world support

Require: first image and aligned depth, text, camera trajectory, adapted video model
Ensure: generated video and retained history
initialize H_1, free-space evidence, and RGB memory in a fixed world frame
for all chunks k **do**
  freeze retained history and retrieve spatial and temporal context
  sample temporary primitives in the upcoming frustum union
  remove observed-space candidates and compute per-view occlusion masks to obtain V_k
  render memory noise from H_k and future noise from $H_k \cup V_k$
  hold noise fixed; denoise future slots conditioned on the known memory
  decode generated RGB and update temporal context
  estimate keyframe depth and validate it against retained geometry
  commit accepted observations with new identities; retain existing identities
  update free-space evidence and retrieval records; discard V_k
end for
return generated video

D EVALUATION METRICS

Rotation error (RotErr ↓). RotErr measures the angular deviation of the recovered camera orientations from the target orientations (Zheng et al., 2024). Let R_t and $\widehat{R}_t$ be target and recovered camera-to-world rotations, each expressed relative to its first matched pose, and let $\mathcal{I}$ contain the matched frame indices. Then

$$\mathrm{RotErr} = \frac{1}{|\mathcal{I}|}\sum_{t\in\mathcal{I}} \theta(R_t^\top \widehat{R}_t), \qquad \theta(R) = \frac{180}{\pi}\arccos\left(\mathrm{clip}\left(\frac{\mathrm{tr}(R)-1}{2}, -1, 1\right)\right). \tag{16}$$

The error is measured in degrees; lower values indicate more accurate camera orientation.

Translation error (TransErr ↓). TransErr measures the discrepancy between target and recovered camera-center trajectories after removing their global origin and scale. Let $\mathbf{c}_t$ and $\widehat{\mathbf{c}}_t$ be camera centers in their respective first-pose coordinate systems, and define $r = \max_{t\in\mathcal{I}} \|\mathbf{c}_t\|_2 + \delta$ and

$\widehat{r} = \max_{t\in\mathcal{I}} \|\widehat{\mathbf{c}}_t\|_2 + \delta$, with a small positive numerical stabilizer δ. The dimensionless error is

$$\mathrm{TransErr} = \frac{1}{|\mathcal{I}|} \sum_{t\in\mathcal{I}} \left\| \frac{\mathbf{c}_t}{r} - \frac{\widehat{\mathbf{c}}_t}{\widehat{r}} \right\|_2 . \tag{17}$$

Relative pose errors (RPE-t ↓, RPE-r ↓). RPE measures the accuracy of camera motion between consecutive frames. Let T_t and $\widehat{T}_t$ denote target and recovered camera-to-world poses after a global similarity alignment, and let $\mathcal{J}$ contain the starting indices of consecutive frame pairs with both poses recovered. Write $\Delta T_t = T_t^{-1} T_{t+1}$, with rotation ΔR_t and translation $\Delta \mathbf{c}_t$, and analogously for $\Delta\widehat{T}_t$. Translation and rotation errors are

$$\begin{aligned} \mathrm{RPE\text{-}t} &= \sqrt{\frac{1}{|\mathcal{J}|} \sum_{t\in\mathcal{J}} \|\Delta\mathbf{c}_t - \Delta\widehat{\mathbf{c}}_t\|_2^2}, \\ \mathrm{RPE\text{-}r} &= \sqrt{\frac{1}{|\mathcal{J}|} \sum_{t\in\mathcal{J}} \theta(\Delta R_t^\top \Delta\widehat{R}_t)^2}. \end{aligned} \tag{18}$$

RPE-t is expressed in the aligned trajectory's translation units, and RPE-r in degrees.

Pose registration coverage (Reg. ↑). Reg. measures the fraction of video frames for which GLOMAP (Pan et al., 2024) successfully recovers a camera pose. For T requested frames,

$$\mathrm{Reg.} = \frac{|\mathcal{I}|}{T}. \tag{19}$$

Higher coverage indicates that poses can be recovered over more of the generated sequence.

Thresholded symmetric epipolar distance (TSED ↑). TSED measures whether feature correspondences between generated views agree with the target camera geometry (Yu et al., 2023). For matched homogeneous image points $\mathbf{p}, \mathbf{q}$ and target fundamental matrix F, the symmetric epipolar distance is

$$d_{\mathrm{sym}}(\mathbf{p},\mathbf{q}) = \frac{d(\mathbf{p}, F^\top\mathbf{q}) + d(\mathbf{q}, F\mathbf{p})}{2}, \qquad d(\mathbf{p},\boldsymbol{\ell}) = \frac{|\boldsymbol{\ell}^\top\mathbf{p}|}{\sqrt{\ell_1^2+\ell_2^2}}. \tag{20}$$

Here d is point-to-line distance in pixels. Let $\mathcal{P}_{\mathrm{epi}}$ contain consecutive frame pairs with nonzero camera baseline, $\mathcal{M}_{ts}$ their feature matches, and m_{ts} their median symmetric epipolar distance ($+\infty$ if no matches are found). The reported score is the fraction of geometrically consistent pairs:

$$\mathrm{TSED} = \frac{1}{|\mathcal{P}_{\mathrm{epi}}|} \sum_{(t,s)\in\mathcal{P}_{\mathrm{epi}}} \mathbf{1}[|\mathcal{M}_{ts}| \geq 10 \ \wedge \ m_{ts} \leq 2 \text{ pixels}] . \tag{21}$$

Higher values indicate stronger cross-view geometric consistency.

E ADDITIONAL VIDEO-GENERATION RESULTS

This section presents additional experimental results of WORLDNOISE.

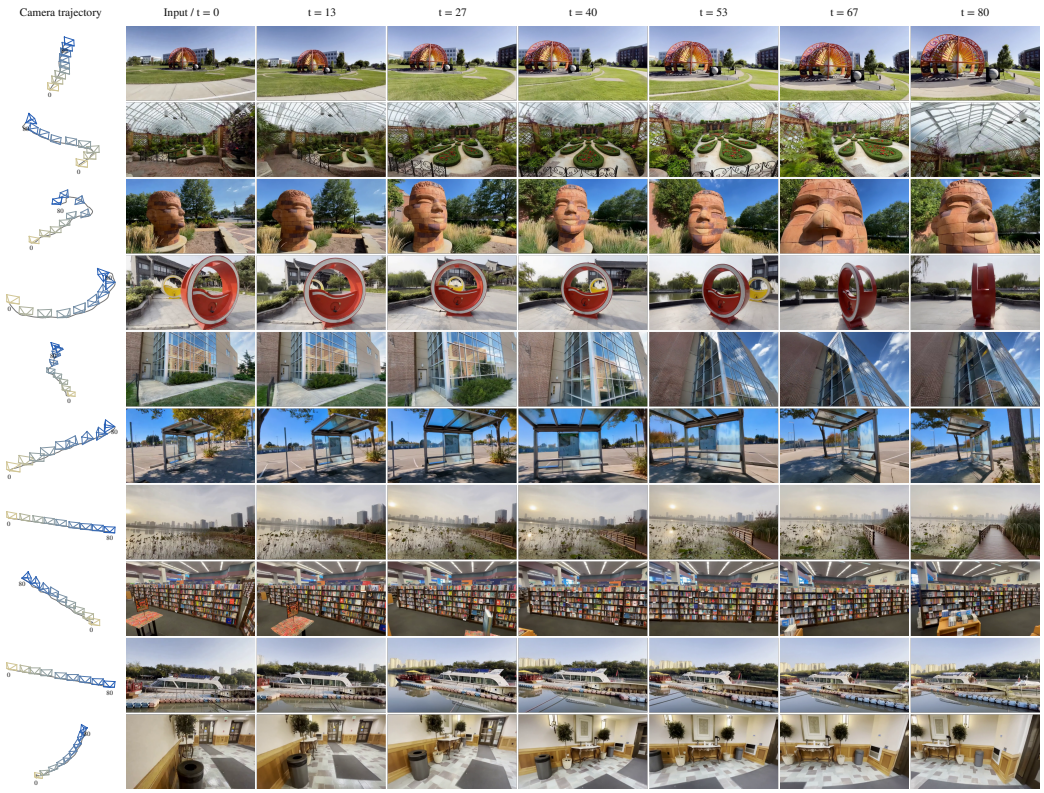

Figure 8: More cases of camera-controlled video generation with WORLDNOISE.

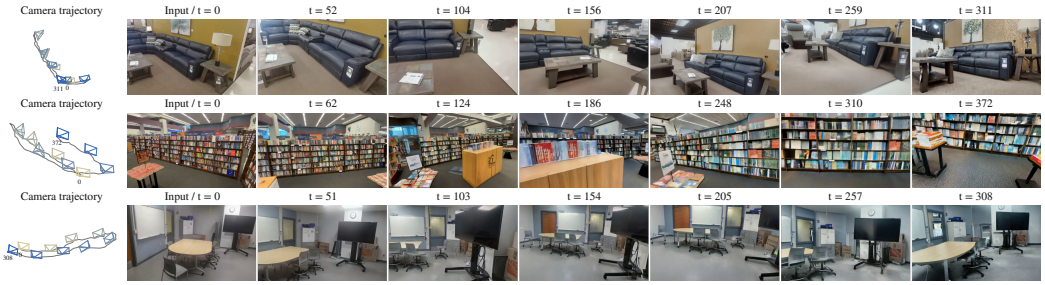

Figure 9: More cases of long-horizon video generation with WORLDNOISE.