VIVAWORLD: LONG-HORIZON VIDEO WORLD MODEL WITH COMPOSITIONAL 3D PRIORS

Ying Jiang [1] **Donglai Xiang** [1] **Anita Hu** [1] **Masha Shugrina** [1]

[1]NVIDIA

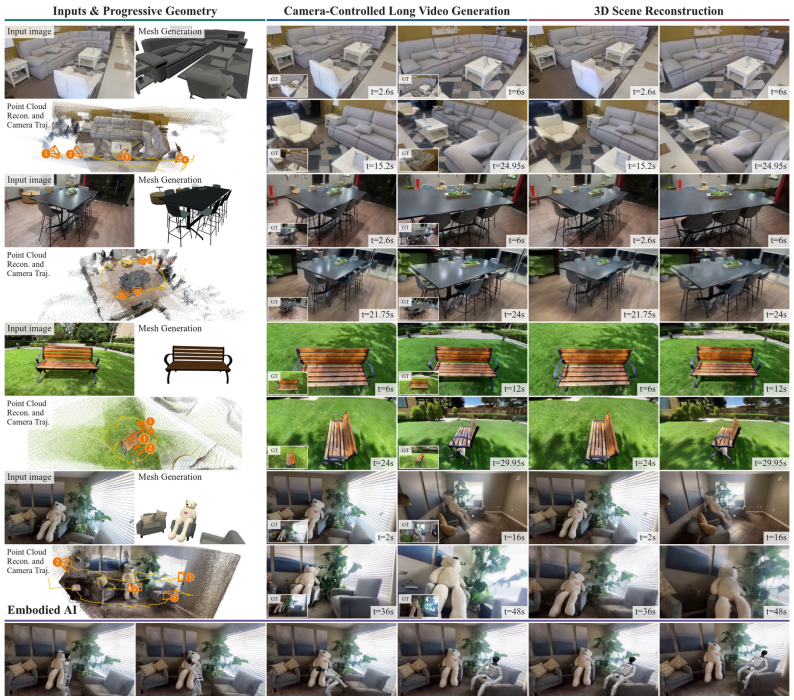

Figure 1: VivaWorld is a novel framework that generates long, geometrically consistent, camera-controlled videos from a single image, enabling further compositional scene reconstruction with disentangled foreground objects and downstream embodied AI robot learning applications.

ABSTRACT

Long-horizon video generation is key to interactive world modeling, physical intelligence, and spatial reasoning, yet maintaining geometric consistency and accurate camera control remains challenging without 3D grounding. We introduce **VivaWorld**, a framework for generating geometry- and view-consistent long videos from an image. We first construct a compositional hybrid 3D representation that combines reconstructed point clouds, which preserve observed scene geometry and appearance, with generated textured meshes, which provide complete object geometry across 360° novel views. We also introduce a connected-region filter removing inconsistent floating components for robust cross-view reconstruction and propose a hybrid camera embedding that combines global and object-anchored Plücker rays to capture both scene-level camera motion and camera-object relationships, and inject it together with the visual guidance into a video generation model via a control adapter. Extensive experiments show **VivaWorld** achieves state-of-the-art visual quality and camera control, while supporting controllable novel-view synthesis, long-horizon scene exploration, and physical AI.

1 INTRODUCTION

Video generation aims to synthesize realistic and spatial-temporally coherent videos from images or text prompts (Singer et al., 2022). Beyond conventional image- or text-conditioned generation, recent video world models are capable of predicting future visual observations conditioned on actions, including camera motions, object trajectories, or agent actions (Shen et al., 2026; Yang et al., 2026; Chen et al., 2026c; Zheng et al., 2026). While early methods primarily focused on generating short video clips conditioned on actions, recent autoregressive video world models (Shen et al., 2026; Chen et al., 2026b) have enabled the synthesis of substantially longer and more controllable videos with richer motions and dynamics. Although individual frames may appear photorealistic, long-horizon video generation remains challenging, as the lack of explicit 3D supervision results in spatial and temporal inconsistencies, including camera drift, geometry distortion, and object deformation, thereby limiting accurate 3D reconstruction and downstream 3D-aware applications.

To enable spatially and temporally consistent controllable video generation, existing approaches leverage explicit 3D representations (Zheng et al., 2026; Chen et al., 2026c; Yang et al., 2026) to reconstruct scenes and render geometrical guidance as 3D memory under target camera viewpoints. Alternatively, recent methods learn implicit geometry priors from 3D latent representations or historical video frames (Huang et al., 2026; Shen et al., 2026). Beyond spatial guidance, recent works (Agarwal et al., 2026; Yang et al., 2026; Shen et al., 2026) have also explored more robust camera representations for temporal consistency. However, existing spatial guidance only makes use of 3D information, e.g., point clouds, extracted from the limited input views, providing weak priors for regions unobservable from those views. Moreover, existing camera embeddings are defined in a global coordinate, making them prone to accumulated errors and long-horizon drift. These limitations raise a question: how can we better leverage 3D grounding to enable consistent long video generation?

We present **VivaWorld**, a framework for generating geometry- and view-consistent 3D-grounded long videos from an image. VivaWorld unifies reconstruction and generative priors through a compositional hybrid representation, providing stronger 3D grounding for long-horizon video generation. Specifically, we leverage feed-forward 3D reconstruction and off-the-shelf 3D generators to reconstruct background point clouds and generate textured foreground meshes. The point clouds preserve observed scene geometry and appearance, while the meshes provide complete object geometry across 360° novel viewpoints. Aggregating point clouds reconstructed from multiple views may introduce inconsistent floating artifacts, so we propose a mesh-seeded connected-region filter that removes disconnected artifacts while preserving valid scene structures. To achieve accurate camera control and mitigate long-horizon drift, we further introduce a compositional camera embedding that combines global Plücker rays with object-anchored rays at fixed registered meshes to explicitly capture camera–object relationships. This design reflects how real-world cameras typically move relative to objects and scenes, such as orbiting around a person or panning across a landscape. By jointly encoding the global camera trajectory and local camera-object offsets, our representation captures both scene-level camera motion and object-relative viewpoint changes. We then introduce a compositional control adapter to inject the hybrid RGB-D controls and compositional camera embeddings into a pretrained video diffusion model, enabling the generation of long-horizon videos with consistent geometry, viewpoints, and camera motion. Extensive experiments demonstrate state-of-the-art performance in visual quality and camera control, while highlighting the versatility of VivaWorld across controllable novel-view synthesis, long-horizon scene exploration, and generative 3D reconstruction. To summarize, our contributions are:

- We propose **VivaWorld**, a novel method for generating view- and geometry-consistent 3D-grounded videos from a single image using a compositional hybrid representation of textured object meshes and background point clouds. We integrate a mesh-seeded filtering strategy into 3D point-cloud reconstruction for improved cross-view robustness.
- We introduce a compositional camera embedding that combines global and object-anchored representations, enabling stable and accurate camera control by capturing both global camera poses and camera–object relationships.

- Extensive experiments demonstrate our method outperforms state-of-the-art approaches in generating accurate, geometry- and view-consistent long videos with precise camera control from a single image, while enabling high-quality compositional 3D reconstruction.

2 Related Work

2.1 Video World Models

Recent advances in diffusion models (Rombach et al., 2022) and transformer architectures (Peebles and Xie, 2023) have sparked significant interest in video world models, which aim to predict future states from sparse multimodal inputs, such as actions, text, images, and other observations (Wang et al., 2026b; Shen et al., 2026; Yang et al., 2026; Agarwal et al., 2026; Yang et al., 2025; Huang et al., 2025b; Zheng et al., 2026). Latent video world models, such as V-JEPA (Assran et al., 2025), predict future latent representations rather than pixels, enabling video understanding, reasoning, and planning. In contrast, generative world models directly synthesize future video frames from multimodal inputs (Wan et al., 2025; Jiang et al., 2025; Wiedemer et al., 2025), enabling high-fidelity and controllable generation. However, they often lack explicit 3D supervision, limiting spatial and temporal consistency under complex or long-range camera motion. To address this, existing approaches leverage explicit 3D representations, such as point clouds (Cao et al., 2025; Zheng et al., 2026; Chen et al., 2026c; Wu et al., 2025b; Yu et al., 2024; 2025), point tracking (Bian et al., 2025) and 3D Gaussian Splatting (Wu et al., 2025a; Yang et al., 2026; De Lutio et al., 2026), to reconstruct scenes and render geometry-aware guidance under target camera viewpoints, providing 3D memory for video generation. Other methods explore historical video frames as complementary guidance (Shen et al., 2026) or learn implicit 3D latent representations that encode scene geometry without explicit reconstruction (Huang et al., 2026). However, existing 3D priors primarily cover regions observed in the input views, leaving unseen regions without reliable geometric guidance and thus with limited 3D consistency. Our goal is to integrate 3D generative priors to cover both observed and unseen regions, enabling spatially and temporally consistent generation.

2.2 Camera Control

To enable accurate camera control, mainstream approaches condition video generation models on either ray maps or explicit camera embeddings. Ray-based methods typically embed a per-pixel 6D representation consisting of ray origins and directions (Popov et al., 2025) or Plücker coordinates (Shen et al., 2026; Yang et al., 2026; He et al., 2024; Wang et al., 2024b; He et al., 2025). However, dense ray maps rely on a world coordinate, potentially limiting generalization (Li et al., 2026), prompting recent works to explore relative ray embeddings (Zhang et al., 2026; Yin et al., 2026). Alternative approaches (Agarwal et al., 2026; Popov et al., 2025; Shin et al., 2026) directly embed camera poses or relative transformations (e.g. to the first or the previous frame) as input. However, none of these representations explicitly model camera–object relationships, causing drift to accumulate over long sequences. Moreover, while point-cloud guidance helps capture camera motion, it may miss coupled drift when the camera and objects shift together, since their pixel-space appearance remains similar. This misalignment propagates to subsequent frames, yielding incorrect camera motion and degraded geometric consistency. Our object-anchored camera embedding anchors camera motion to scene objects to explicitly model their relationships, while hybrid mesh and point-cloud guidance mitigates both camera and object drift, enabling accurate camera control and geometrically consistent long-video generation.

2.3 3D Generation and Reconstruction

Prior work reconstructs or generates 3D shapes using optimization-based or feed-forward methods across representations such as point clouds, meshes, 3DGS, and NeRFs (Go et al., 2025; Schmid et al., 2026; Held et al., 2026; Xia et al., 2026; Bahmani et al., 2026; Mildenhall et al., 2021; Kerbl et al., 2023). While classical per-instance methods, including Structure-from-Motion (SfM) (Schonberger and Frahm, 2016), Multi-View Stereo (MVS) (Goesele et al., 2006), and optimization-based reconstruction (Mildenhall et al., 2021; Kerbl et al., 2023) achieve high reconstruction quality, they incur substantial computational costs. In contrast, recent feed-forward 3D foundation models (Wang

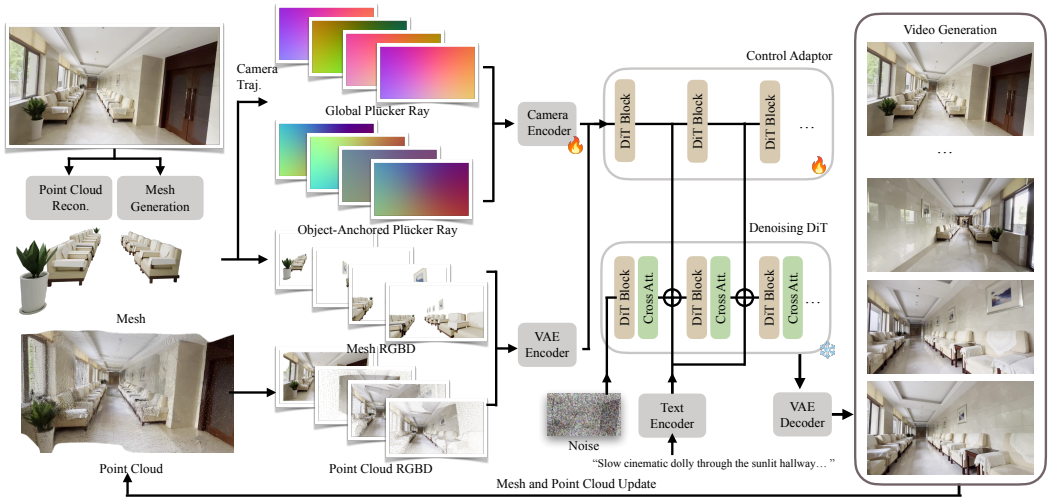

Figure 2: **Pipeline Overview.** Given an image, **VivaWorld** renders extracted point clouds and posed meshes along the camera trajectory as visual guidance, and encodes camera motion with global and object-anchored Plücker rays to capture both scene-level motion and camera-object relationships. A compositional control adapter injects those signals into a pretrained video diffusion model, and generated frames update visual guidance to condition the next chunk for long-horizon generation.

et al., 2025a; 2026c;a; Lin et al., 2025; Ren et al., 2026; Lu et al., 2025; Zhao et al., 2026; Ren et al., 2024; Wang et al., 2024a; 2025b) efficiently infer scene geometry in a single pass, making them well-suited for providing 3D guidance to video generation models. 3D reconstruction and generation methods have different focuses. Reconstruction (Wang et al., 2025a; 2026c;a) provides accurate geometry only in observed regions but offer limited priors for unseen regions; existing work on video generation have explored 3D grounding using this type of priors more. Conversely, 3D generation models (Lai et al., 2025; Chen et al., 2026a) provide complete shape priors for arbitrary viewpoints but may not faithfully align with observations. We use the high fidelity of reconstruction and the completeness of generation as complementary hybrid 3D priors for long video generation.

3 Method

Given a reference image $\mathbf{I}_{\text{ref}}$, a text prompt, and a camera trajectory $\mathcal{C} = \{(K_t, E_t)\}_{t=1}^{T}$ with intrinsics K_t and camera-to-world poses $E_t \in \text{SE}(3)$, our goal is to generate a spatially and temporally consistent video $\mathcal{V} = \{\mathbf{I}_t\}_{t=1}^{T}$ that follows $\mathcal{C}$. As illustrated in Fig. 2, we construct a compositional geometric representation $\mathcal{W}_{\text{geo}} = \{\mathcal{P}, \mathcal{M}\}$ of a background point cloud $\mathcal{P}$ and posed foreground meshes $\mathcal{M}$ (Sec. 3.1) and render it along $\mathcal{C}$ to obtain RGB-D conditions $\mathbf{c}_{\text{rgbd}}$. We further introduce an object-anchored camera representation $\mathbf{c}_{\text{cam}}$ encoding both global and object-relative camera motion (Sec. 3.2). Both conditions are injected into a pretrained video diffusion transformer through lightweight adapters (Sec. 3.3). For long-horizon generation, we divide $\mathcal{C}$ into consecutive chunks $\mathcal{C}^{(q)}$ of up to L poses and generate the corresponding video chunks $\mathcal{V}^{(q)}$ autoregressively, each conditioned on $\mathbf{c}_{\text{rgbd}}^{(q)}$, $\mathbf{c}_{\text{cam}}^{(q)}$, and previously generated frames as temporal context, while newly generated frames update $\mathcal{P}$ for subsequent chunks.

3.1 Compositional Geometric Control

To preserve spatial consistency, we use a hybrid point-cloud and mesh representation $\mathcal{W}$ as geometric conditioning. We reconstruct a background point cloud $\mathcal{P}$ from observed regions and generate complete foreground meshes $\mathcal{M}$. Rendering $\mathcal{W}$ along the target camera trajectory $\mathcal{C}$ produces RGB-D conditions $\mathbf{c}_{\text{rgbd}}$ for consistent video generation.

Foreground Object Mesh Generation. To obtain object geometry that is both complete and aligned with the observations, we decompose posed-object generation into geometry generation and scene-

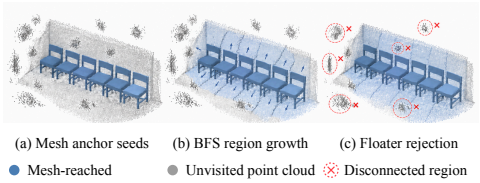

Figure 3: **Mesh-Seeded Point Cloud Filtering.** (a) Object meshes (blue) provide seeds. (b) BFS propagates through occupied voxels, leaving disconnected floating clusters unreached. (c) Unvisited components (red) are removed.

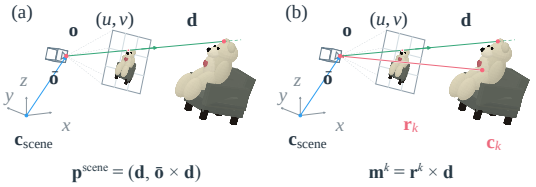

Figure 4: **Compositional Camera Representation.** (a) The global ray $\mathbf{p}^{\text{scene}}$ uses the camera offset from the scene center. (b) Each object-anchored moment $\mathbf{m}^k$ is the cross product of the object-relative offset $\mathbf{r}^k$ from object center $\mathbf{c}^k$ and the ray direction $\mathbf{d}$.

space registration. We segment objects in $\mathbf{I}_{\text{ref}}$ with SAM 3 (Carion et al., 2026) and generate a mesh $\widehat{\mathcal{M}}^k$ for each object k using RecGen (Zadaianchuk et al., 2026) or GPT-6 Astra (OpenAI, 2026). We then align $\widehat{\mathcal{M}}^k$ to its observed points $\mathcal{P}^k_{\text{obj}}$, extracted from the Pi3 (Wang et al., 2026c) point cloud using the object's mask, via Iterative Closest Point (ICP) registration with per-axis scaling, estimating $T^k(\mathbf{x}) = R^k D^k \mathbf{x} + \mathbf{t}^k$ with rotation $R^k \in \mathrm{SO}(3)$, translation $\mathbf{t}^k \in \mathbb{R}^3$, and anisotropic scale $D^k = \mathrm{diag}(s^k_x, s^k_y, s^k_z)$. The registered meshes form $\mathcal{M} = \{T^k(\widehat{\mathcal{M}}^k)\}_k$.

Background Point Cloud Reconstruction. For each chunk, we use Pi3 (Wang et al., 2026c) to predict dense metric depth and unproject it using the camera intrinsics and poses estimated by ViPE (Huang et al., 2025a), producing a local point cloud anchored at the chunk's first frame. Aggregating these local reconstructions yields a global point cloud $\mathcal{P}_{\text{raw}}$. However, inconsistent depth estimates across chunks may introduce floating artifacts in mid-air that are disconnected from the main scene. To remove them, we propose a mesh-seeded connected-region filter (Fig. 3) using foreground meshes as geometric anchors. We voxelize $\mathcal{P}_{\text{raw}}$ and $\mathcal{M}$, then flood-fill from mesh-occupied voxels through neighboring occupied voxels, and retain only points in reached voxels, removing all components disconnected from the registered meshes. The full algorithm is in Appendix A.1.

Compositional Visual Control. Given $\mathcal{W}_{\text{geo}} = \{\mathcal{P}, \mathcal{M}\}$, we render $\mathcal{P}$ and $\mathcal{M}$ separately along each chunk $\mathcal{C}^{(q)}$, yielding point-cloud and mesh RGB-D renderings $I^{\mathcal{P}}$ and $I^{\mathcal{M}}$ that form $\mathbf{c}^{(q)}_{\text{rgbd}}$. We further derive a binary visual-guidance mask $\mathbf{b}$ from the union of the point-cloud and mesh rendering masks, indicating where either rendering provides visual guidance. The point cloud preserves observed appearance, while the meshes provide complete object geometry and clearer camera-motion cues through coherent view-dependent projections.

3.2 Compositional Object-Oriented Camera Control

Inspired by the observation that cameras typically move relative to objects and scenes, we propose a compositional camera embedding $\mathbf{c}_{\text{cam}}$ combining the camera's global pose with its relative position to foreground objects to capture scene-level camera motion and object-centric viewpoint changes.

Scene Camera Embedding. For frame t with camera-to-world pose $E_t = [U_t \mid \mathbf{o}_t]$, where U_t is the rotation and $\mathbf{o}_t$ the camera center, and intrinsics K_t, each pixel (u, v) has world-space ray direction $\mathbf{d}_{t,u,v} = \mathrm{normalize}(U_t K_t^{-1} \widetilde{\mathbf{u}})$ with $\widetilde{\mathbf{u}} = (u, v, 1)^\top$. Its Plücker coordinate is

$$\mathbf{p}^{\text{scene}}_{t,u,v} = (\mathbf{d}_{t,u,v}, \bar{\mathbf{o}}_t \times \mathbf{d}_{t,u,v}), \qquad \bar{\mathbf{o}}_t = \frac{\mathbf{o}_t - \mathbf{c}_{\text{scene}}}{s_{\text{scene}}}, \tag{1}$$

where the scene center $\mathbf{c}_{\text{scene}}$ and scale s_{scene} are computed from the camera and foreground-object centers. Stacking over all frames and pixels yields $\mathbf{p}^{\text{scene}} \in \mathbb{R}^{T \times 6 \times H \times W}$.

Object-Anchored Camera Embedding. Global Plücker rays do not explicitly capture camera motion relative to individual objects. We therefore use the registered meshes reconstructed from $\mathbf{I}_{\text{ref}}$ as fixed object anchors with centers $\{\mathbf{c}^k\}_{k=1}^N$ in world coordinates. For each object k, we shift the ray origin to its center and compute the object-anchored moment as

$$\mathbf{m}^k_{t,u,v} = \frac{\mathbf{o}_t - \mathbf{c}^k}{s_{\text{scene}}} \times \mathbf{d}_{t,u,v}, \qquad \mathbf{c}_{\text{cam}} = \left[\mathbf{p}^{\text{scene}} \middle\| \mathbf{m}^1 \middle\| \cdots \middle\| \mathbf{m}^N\right]. \tag{2}$$

Since $\mathbf{d}_{t,u,v}$ is shared across anchors, only the moment is added for each object, where $\|$ denotes channel-wise concatenation. The resulting tensor $\mathbf{c}_{\mathrm{cam}} \in \mathbb{R}^{T\times(6+3N)\times H\times W}$ jointly encodes global camera motion and object-relative viewpoint changes.

3.3 Model Design

Model Architecture. We adopt the pretrained Wan 2.1 T2V-14B model as our frozen video diffusion backbone and build our compositional control adapter on VACE (Jiang et al., 2025). Given a target video chunk, the frozen Wan VAE encodes it into a latent $\mathbf{z}_0$, whose noisy version $\mathbf{z}_\tau$ at diffusion timestep τ is patchified into video tokens $\mathbf{x}_\tau$. The point-cloud and mesh RGB-D renderings $I^{\mathcal{P}}$ and $I^{\mathcal{M}}$ are encoded by the frozen Wan VAE, with RGB and depth encoded separately as $\mathcal{E}_{\mathrm{RGBD}}(I) = [\mathcal{E}_{\mathrm{VAE}}(I_{\mathrm{rgb}})\|\mathcal{E}_{\mathrm{VAE}}(I_{\mathrm{depth}})]$, while a lightweight spatiotemporal convolutional encoder Φ_{cam} encodes the mask $\mathbf{b}$ and camera embedding $\mathbf{c}_{\mathrm{cam}}$. The resulting control latent is

$$\mathbf{z}_{\mathrm{ctrl}} = \left[\mathcal{E}_{\mathrm{RGBD}}(I^{\mathcal{P}})\middle\|\mathcal{E}_{\mathrm{RGBD}}(I^{\mathcal{M}})\middle\|\Phi_{\mathrm{cam}}([\mathbf{b}\|\mathbf{c}_{\mathrm{cam}}])\right]. \tag{3}$$

The control latent is patchified into control tokens aligned one-to-one with $\mathbf{x}_\tau$ and processed by a transformer branch that injects residual features into selected blocks of the frozen Wan-DiT, sharing its timestep and text conditioning. We train only the adapter and camera encoder using the backbone's flow-matching objective. More architecture details are provided in Appendix A.2.

Supporting Single- and Multi-Image Inputs. For a single input image, we initialize the point cloud from $\mathbf{I}_{\mathrm{ref}}$ and update it causally using newly generated chunk-boundary frames. For sparse multi-image inputs, we use each image as the anchor frame of a corresponding video chunk and jointly reconstruct a global point cloud from all available views. During training, we simulate both settings (see Appendix A.3 for training and data curation details).

4 Experiments

4.1 Experimental Setup

Implementation Details. We build on VACE (Jiang et al., 2025) with the frozen Wan 2.1-14B backbone (Wan et al., 2025) and train only the compositional control adapter (3.05B parameters) and camera encoder on DL3DV-10K (Ling et al., 2024). Data and more details are in Appendix.

Baselines and Evaluation Protocol. We evaluate our method from two perspectives: video generation and 3D scene reconstruction. For video generation, we compare against six representative camera-controlled methods: the static-scene video generation baselines AnyRecon (Chen et al., 2026c), Gen3C (Ren et al., 2025), Voyager (Huang et al., 2025b), and Lyra 2.0 (Shen et al., 2026), and the dynamic-scene video generation baselines VerseCrafter (Zheng et al., 2026) and NeoVerse (Yang et al., 2026). Among them, AnyRecon, Gen3C, and Lyra 2.0 support world-consistent long-video generation. We evaluate on 20 scenes ($\sim$ 200 video clips) from DL3DV-10K (Ling et al., 2024) test set and Tanks and Temples (Knapitsch et al., 2017) to assess in-domain performance and cross-dataset generalization. We evaluate both single-chunk and long-video generation at 832×480 resolution. In the single-chunk setting, we explore ground-truth initial frames from chunks as the input and evaluate only the frames of each chunk common to all methods since chunk lengths differ. For long-video generation, we evaluate autoregressive predictions over the first 1,000 frames of each DL3DV scene and the first 240 frames of each Tanks and Temples scene. We report PSNR, SSIM, and LPIPS for visual fidelity and consistency and adapted WorldScore (Duan et al., 2025) metrics for style and camera consistency. We also report rotation and unit-normalized translation errors (RotErr and TransErr) for camera-control accuracy. See more details in the appendix for the style metric and the camera metrics. For the quantitative study, we reconstruct 3D Gaussian splats from the generated videos using Depth Anything 3 (Lin et al., 2025) and evaluate the rendered views against the ground-truth images using SSIM and LPIPS. For the qualitative study, we additionally explore 3DGS optimization for scene reoncstruction from generative videos.

Table 1: **Video Generation Scores.** Baselines degrade sharply from single chunks to long videos, while VivaWorld achieves the best visual fidelity in both and the best long-video camera control.

	Single-Chunk Generation					Long-Video Generation				
Methods	PSNR↑	SSIM↑	LPIPS↓	Style↑	Cam.↑	PSNR↑	SSIM↑	LPIPS↓	Style↑	Cam.↑
AnyRecon	18.54	0.436	0.285	98.67	94.75	12.64	0.361	0.600	96.37	27.37
Gen3C	19.92	0.513	0.254	98.09	91.61	12.71	0.333	0.609	**96.99**	62.56
Voyager	18.60	0.446	0.345	98.62	92.57	8.622	0.213	0.796	91.96	55.58
Lyra2	20.01	0.528	0.227	98.74	**96.34**	12.85	0.343	0.582	96.82	65.09
VerseCrafter	17.24	0.431	0.292	97.45	85.53	11.00	0.294	0.691	95.48	27.76
NeoVerse	17.70	0.415	0.370	**99.02**	96.22	10.42	0.243	0.704	93.67	49.09
Ours	**21.24**	**0.570**	**0.173**	98.65	95.22	**13.92**	**0.376**	**0.529**	96.14	**79.51**

Table 2: **Camera Control and 3D Scene Generation.** VivaWorld achieves the lowest long-video camera errors and the most accurate 3D reconstructions.

	Camera Control				3D Scene Generation			
	Single-Chunk Gen.		Long-Video Gen.		Single-Chunk Gen.		Long-Video Gen.	
	RotErr↓	TransErr↓	RotErr↓	TransErr↓	SSIM↑	LPIPS↓	SSIM↑	LPIPS↓
AnyRecon	0.375	0.036	28.47	0.212	0.461	0.449	0.444	0.673
Gen3C	0.655	0.056	14.30	0.121	0.472	0.448	0.418	0.676
Voyager	0.452	0.051	14.91	0.191	0.461	0.492	0.350	0.802
Lyra2	**0.215**	0.025	9.022	0.098	0.472	0.439	0.423	0.658
VerseCrafter	0.604	0.107	28.91	0.257	0.451	0.475	0.403	0.763
NeoVerse	0.400	**0.023**	22.14	0.189	0.470	0.460	0.415	0.781
Ours	0.308	0.033	**6.908**	**0.060**	**0.480**	**0.417**	**0.447**	**0.598**

4.2 Quantitative Evaluation

Tab. 1 compares single-chunk generation and long-video generation over 1,000 frames. Most baselines perform well on short clips but degrade substantially over longer horizons. NeoVerse achieves the highest short-clip style score, second best camera score and Voyager remains competitive on short clips, but both suffer large declines in visual quality over long horizons, with Voyager degrading a lot. VerseCrafter and AnyRecon exhibit the weakest long-video camera control, while Gen3C maintains strong style consistency but accumulates camera error. Lyra2 is the strongest baseline overall. Our method achieves the best PSNR, SSIM, and LPIPS in both settings and the best long-video camera-control score, while Gen3C achieves the highest long-video style score. As shown in Tab. 2, our method reduces long-video rotation and translation errors by 23.4% and 38.8%, respectively, compared with Lyra2, demonstrating that our hybrid 3D guidance and object-anchored camera embedding effectively mitigate accumulated drift. For downstream 3D scene generation, our method achieves the best SSIM and LPIPS in both settings, whereas all baselines show reduced reconstruction quality as video length increases. By incorporating object anchors into both visual guidance and camera embeddings, our method preserves stronger cross-frame geometric consistency, resulting in more accurate and coherent 3D reconstructions.

4.3 Qualitative Evaluation

Fig. 5 presents qualitative comparisons among video generation methods. NeoVerse, VerseCrafter, and Voyager perform well over short horizons but degrade over longer ones: NeoVerse propagates reconstruction artifacts under large camera motion, while VerseCrafter and Voyager lose global spatial consistency, producing incorrect object counts (e.g., yachts or chairs). Lyra 2.0, Gen3C, and AnyRecon maintain better long-horizon quality but still degrade: Lyra 2.0 and Gen3C accumulate camera drift and lose texture detail, and AnyRecon becomes increasingly blurry. In contrast, our method accurately follows the target trajectory while preserving appearance and geometric consis-



Figure 5: **Video Generation.** Baselines show color artifacts, distortion, or incorrect object counts, while **VivaWorld** preserves the ground-truth layout and appearance.

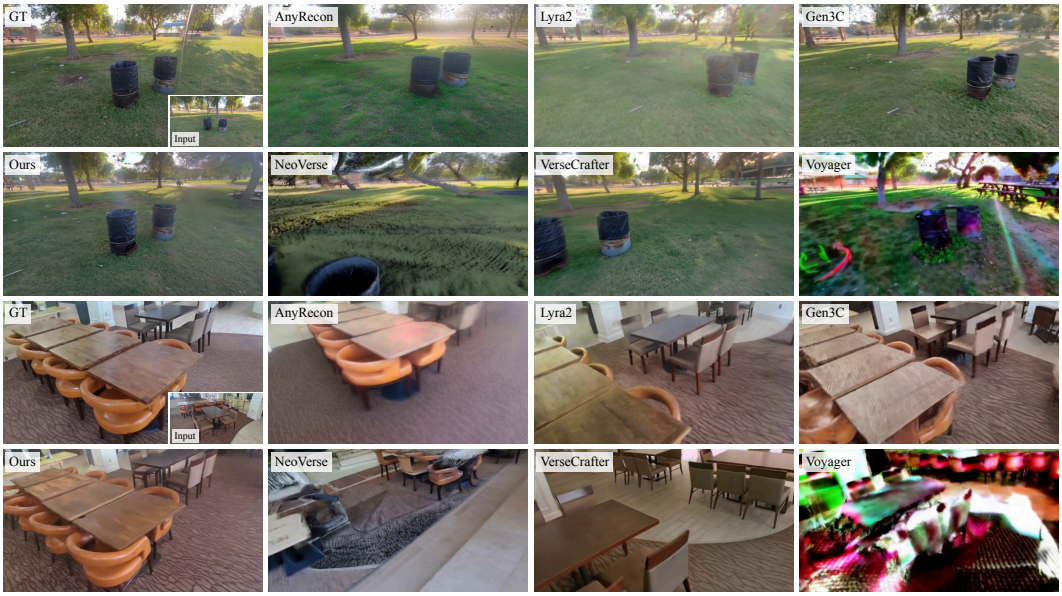

Figure 6: **3D Reconstruction.** We reconstruct 3D scenes from generated videos with per-scene 3DGS optimization. Baselines produce blurry geometry, while ours yields a clean reconstruction.

tency throughout. For downstream 3D reconstruction from generated videos (Fig. 6), we evaluate using optimization-based reconstruction for 3000 steps. For baseline methods, inconsistent object positions across generated frames lead to duplicated or blurry geometry, and objects can even take on a different shape from the input frame. In contrast, our method produces more accurate and coherent reconstructions. Additional results are provided in the appendix and supplementary video.

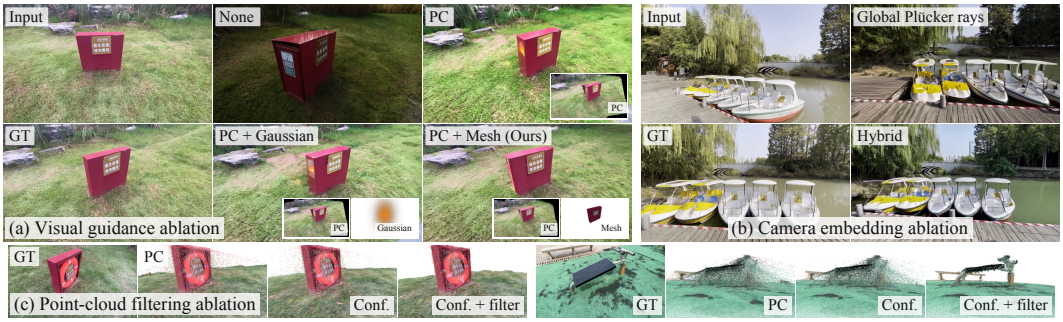

Figure 7: **Ablation.** Ablation studies showing our full configuration outperforms alternative settings.

Table 3: **Ablation Study.** The ablation study indicates that exploring hybrid visual guidance and composition yields consistent gains, achieving better performance than the other settings.

	Single-Chunk			Long-Video		
Guidance Signal	PSNR ↑	LPIPS ↓	Cam ↑	PSNR ↑	LPIPS ↓	Cam ↑
None	11.47	0.490	60.35	10.03	0.700	3.17
PC	20.00	0.174	89.66	12.41	0.584	71.34
PC + Gaussian	20.76	0.168	83.94	13.60	0.550	79.95
PC + Mesh + Global Embedding	20.49	0.149	89.84	13.56	0.512	78.32
Ours (PC + Mesh + Hybrid Embedding)	**21.62**	**0.145**	**90.69**	**14.30**	**0.488**	**80.97**

4.4 Ablation

We ablate the key components of the proposed framework in the DL3DV test set.

3D Compositional Guidance. We compare the hybrid mesh and point-cloud guidance against no visual guidance ("None"), point clouds alone ("PC"), and point clouds augmented with object-level 3D Gaussian proxies ("PC+Gaussian") as in (Zheng et al., 2026) (Tab. 3). All guided variants outperform the unguided one, which suffers substantial camera drift and deviation from the target trajectory in long-video generation. Point clouds improve camera alignment by providing global 3D memory, and Gaussian proxies further aid camera control by making camera motion easier to perceive, particularly over repetitive background textures such as grass. Integrated point cloud guidance with meshes improves not only camera consistency but also image quality via stronger geometry priors.

Camera Embedding. Tab. 3 compares different Plücker-ray embeddings. In Fig. 7, hybrid embedding improves visual quality and controllability.

Filtered Point Cloud. In Fig. 7(c), we compare the reconstructed point cloud without confidence score and connection filtering, the point cloud after confidence score filtering, and our point cloud guidance using both confidence scores and the proposed connectivity-based filtering. Confidence-based filtering removes some artifacts but leaves substantial floating points, particularly near object boundaries. Our method removes floating points near foreground objects while suppressing background noise, providing cleaner visual guidance.



Figure 8: **Agentic Mesh and Video Generation.** Given an image and an initial mesh, our method generates novel views that provide visual references for refining agentic generation.



Figure 9: **Video Interpolation.** Our method generates intermediate frames from keyframes.

4.5 Applications

Agentic Video Generation. We support agentic 3D objects as compositional guidance, enabling iterative 3D-video co-refinement. In Fig. 8, a GPT-generated mesh guides video generation along the target camera trajectory, and the resulting video refines the mesh quality and spatial arrangement.

Video Interpolation. We also support chunk-wise high-quality video interpolation between sparse input views, using a fixed global point cloud and object meshes as geometric guidance (Fig. 9).

Humanoid Robot Learning. The generated videos enable compositional 3D reconstruction, supporting humanoid robot learning of walking and sitting tasks (Fig. 1).

5 Conclusion, Limitation and Future Work

We presented **VivaWorld**, a novel framework for generating long, 3D-grounded videos with spatial consistency and accurate camera control from a single image. By integrating point-cloud reconstruction priors with generative mesh priors as hybrid visual guidance and introducing compositional camera conditioning, VivaWorld produces spatially coherent video frames over extended trajectories. However, the current pipeline does not fully prevent color drift during long-horizon generation. Future work could explore explicit appearance-consistency constraints or color alignment strategies to preserve consistent colors throughout the long generation video.

References

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In *International Conference on Learning Representations*, volume 2026, pages 82850–82880, 2026.

Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. Gs-

dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. *arXiv preprint arXiv:2501.02690*, 2025.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–12, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pages 138846–138923, 2026.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7220–7232, 2026a.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026b.

Yutian Chen, Shi Guo, Renbiao Jin, Tianshuo Yang, Xin Cai, Yawen Luo, Mingxin Yang, Mulin Yu, Linning Xu, and Tianfan Xue. Anyrecon: Arbitrary-view 3d reconstruction with video diffusion model. *arXiv preprint arXiv:2604.19747*, 2026c.

Riccardo De Lutio, Tobias Fischer, Yen-Yu Chang, Yuxuan Zhang, Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Katarína Tóthová, Zan Gojcic, and Haithem Turki. Artifixer: Enhancing and extending 3d reconstruction with auto-regressive diffusion models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2026.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 27713–27724, 2025.

Hyojun Go, Dominik Narnhofer, Goutam Bhat, Prune Truong, Federico Tombari, and Konrad Schindler. Vist3a: Text-to-3d by stitching a multi-view reconstruction network to a video generator. *arXiv e-prints*, pages arXiv–2510, 2025.

Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2402–2409. IEEE, 2006.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 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*, pages 13416–13426, 2025.

Jan Held, Renaud Vandeghen, Adrien Deliege, Abdullah Hamdi, Silvio Giancola, Daniel Rebain, Anthony Cioppa, Bernard Ghanem, Andrea Vedaldi, Andrea Tagliasacchi, et al. Triangle splatting for real-time radiance field rendering. In *2026 International Conference on 3D Vision (3DV)*, pages 1248–1257. IEEE, 2026.

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*, 2025a.

Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25358–25369, 2026.

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

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

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

Zeqiang Lai, Yunfei Zhao, Haolin Liu, Zibo Zhao, Qingxiang Lin, Huiwen Shi, Xianghui Yang, Mingxin Yang, Shuhui Yang, Yifei Feng, et al. Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. *arXiv preprint arXiv:2506.16504*, 2025.

Ruilong Li, Brent Yi, Junchen Liu, Hang Gao, Yi Ma, and Angjoo Kanazawa. Cameras as relative positional encoding. *Advances in Neural Information Processing Systems*, 38:15984–16009, 2026.

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.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024.

Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 27272–27283. IEEE, 2025.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

OpenAI. GPT-6 astra: A new generation of intelligence, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

Stefan Popov, Amit Raj, Michael Krainin, Yuanzhen Li, William T Freeman, and Michael Rubinstein. Camctrl3d: Single-image scene exploration with precise 3d camera control. In *2025 International Conference on 3D Vision (3DV)*, pages 649–658. IEEE, 2025.

Jiawei Ren, Michal Jan Tyszkiewicz, Jiahui Huang, and Zan Gojcic. Tokengs: Decoupling 3d gaussian prediction from pixels with learnable tokens. *arXiv preprint arXiv:2604.15239*, 2026.

Xuanchi Ren, Yifan Lu, Hanxue Liang, Zhangjie Wu, Huan Ling, Mike Chen, Sanja Fidler, Francis Williams, and Jiahui Huang. Scube: Instant large-scale scene reconstruction using voxsplats. *Advances in Neural Information Processing Systems*, 37:97670–97698, 2024.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Katharina Schmid, Nicolas von Lützow, Jozef Hladkỳ, Angela Dai, and Matthias Nießner. Genrecon: Bridging generative priors for multi-view 3d scene reconstruction. *arXiv preprint arXiv:2605.23888*, 2026.

Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

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

Seunghyun Shin, Jifei Song, Wooseok Jeon, Hae-Gon Jeon, and Jiankang Deng. Trimotion: Modality-agnostic camera control for video generation. *arXiv preprint arXiv:2606.20774*, 2026.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

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.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5294–5306. IEEE, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. Vggt-omega. *arXiv preprint arXiv:2605.15195*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10510–10522. IEEE, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709. IEEE, 2024a.

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

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations*, volume 2026, pages 10481–10497, 2026c.

Yuelei Wang, Jian Zhang, Pengtao Jiang, Hao Zhang, Jinwei Chen, and Bo Li. Cpa: Camera-pose-awareness diffusion transformer for video generation. *arXiv preprint arXiv:2412.01429*, 2024b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26024–26035. IEEE, 2025a.

Tong Wu, Shuai Yang, Ryan Po, Yinghao Xu, Ziwei Liu, Dahua Lin, and Gordon Wetzstein. Video world models with long-term spatial memory. *arXiv preprint arXiv:2506.05284*, 2025b.

Hongchi Xia, Xuan Li, Zhaoshuo Li, Qianli Ma, Jiashu Xu, Ming-Yu Liu, Yin Cui, Tsung-Yi Lin, Wei-Chiu Ma, Shenlong Wang, et al. Sage: Scalable agentic 3d scene generation for embodied ai. *arXiv preprint arXiv:2602.10116*, 2026.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, et al. Grail: Generating humanoid loco-manipulation from 3d assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026.

Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14662–14672. IEEE, 2024.

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

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Shan Ying, and Kai Han. Raype: Ray-space positional encoding for 3d-aware video generation. *arXiv preprint arXiv:2606.27345*, 2026.

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

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.

Andrii Zadaianchuk, Leonardo Barcellona, Lennard Schuenemann, Christian Gumbsch, Zehao Wang, Muhammad Zubair Irshad, Fabien Despinoy, Rahaf Aljundi, Stratis Gavves, and Sergey Zakharov. Reconstruction by generation: 3d multi-object scene reconstruction from sparse observations. *arXiv preprint arXiv:2604.27106*, 2026.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo 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*, pages 38027–38037, 2026.

Qitao Zhao, Hao Tan, Qianqian Wang, Sai Bi, Kai Zhang, Kalyan Sunkavalli, Shubham Tulsiani, and Hanwen Jiang. E-rayzer: Self-supervised 3d reconstruction as spatial visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7525–7535, 2026.

Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. Versecrafter: Dynamic realistic video world model with 4d geometric control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 40277–40290, 2026.

A METHOD & TRAINING DETAILS

A.1 MESH-SEEDED POINT CLOUD FILTERING

Given the raw point cloud $\mathcal{P}_{\text{raw}}$ and foreground meshes $\mathcal{M}$, we voxelize both into a grid of resolution $R \times R \times R$. Let $\mathcal{O}_R = \text{Vox}_R(\mathcal{P}_{\text{raw}}) \cup \text{Vox}_R(\mathcal{M})$ denote all occupied voxels and $\mathcal{S}_R = \text{Vox}_R(\mathcal{M})$ the mesh-occupied seed voxels. Starting from $\mathcal{S}_R$, we perform multi-source flood filling over $\mathcal{O}_R$. At each step, an occupied voxel $\mathbf{v}$ expands to neighboring occupied voxels $\mathbf{u}$ satisfying $0 < \|\mathbf{u} - \mathbf{v}\|_\infty \leq r$. Denoting all visited voxels by $\mathcal{R}_r$, we obtain the filtered point cloud as

$$\mathcal{P} = \{\mathbf{p} \in \mathcal{P}_{\text{raw}} \mid \text{Vox}_R(\mathbf{p}) \in \mathcal{R}_r\}, \tag{4}$$

where $\text{Vox}_R(\mathbf{p})$ returns the voxel containing $\mathbf{p}$. This operation retains scene geometry connected to at least one registered mesh and removes all unvisited components as floaters.

We evaluate conditioning quality across five cases by comparing renderings of progressively reconstructed point clouds with ground-truth images. Table 4 reports mean PSNR gains over unfiltered point clouds. Confidence filtering at a threshold of 0.01 improves conditioning quality by suppressing reconstruction noise, while mesh-seeded filtering achieves the largest gain with a $128\times128\times128$ voxel grid and $r = 1$. We nevertheless choose $r = 2$, which achieves a nearly identical gain while allowing connections across a one-voxel gap, providing greater tolerance to fragmented geometry.

Table 4: **Mesh-seeded Filtering Ablation.** Mean conditioning PSNR gains (dB) over unfiltered point clouds across five cases. Grid size denotes the voxel resolution along each axis, and r denotes the connectivity radius.

Confidence	Grid 32^3		Grid 64^3		Grid 128^3	
≥ 0.01	$r = 1$	$r = 2$	$r = 1$	$r = 2$	$r = 1$	$r = 2$ (Ours)
$+0.0512$	$+0.0028$	$+0.0080$	$+0.0532$	$+0.0531$	$\mathbf{+0.1402}$	$\mathbf{+0.1387}$

In addition to using a connectivity radius of $r = 2$, we use multiple object meshes derived from SAM's "everything" mode as seeds to preserve valid regions disconnected by occlusion or sparse sampling, increasing coverage across the scene. We acknowledge that mesh-seeded filtering parameters, including voxel size and connectivity radius, may require adjustment for complex scenes or inaccurate reconstructions. Each filtering run takes approximately one second on average, making it practical to evaluate several parameter settings. Selecting the best-performing setting for each case yields a mean PSNR gain of $+0.2421$ dB across these five cases.

A.2 CAMERA ENCODER

The camera encoder Φ_{cam} is a lightweight single-layer 3D CNN that takes the concatenation of the visual-guidance mask $\mathbf{b}$ and the camera embedding $\mathbf{c}_{\text{cam}}$ (16 channels for $N = 3$ anchors) and applies one Conv3D with kernel size and stride $(4, 8, 8)$, followed by GroupNorm and SiLU, producing a 64-channel feature aligned with the $4\times$ temporal and $8\times$ spatial compression of the Wan VAE latent. Since the camera embedding is already a dense, pixel-aligned geometric signal, this single layer is sufficient, and the encoder contains only 0.26M parameters. By default, we consider three object candidates as anchors. If mesh generation produces fewer than three meshes, we pad the set to three and if it produces more than three, we select the three 3D objects most similar to the segmented image, based on image–object similarity.

A.3 GEOMETRIC CONDITIONING

To support both single- and multi-image inputs, we curate two complementary point-cloud conditions. For multi-image conditioning, we uniformly sample 20 frames across the video, use each as the anchor of a corresponding chunk, and jointly reconstruct a global point cloud $\mathcal{P}^{\text{global}}$ shared across all chunks. For single-image autoregressive generation, we construct a causal point cloud $\mathcal{P}^{(q)}_{\text{causal}}$ for each chunk q using only $\mathbf{I}_{\text{ref}}$ and the anchor frames of preceding chunks, so the first

chunk uses only $\mathbf{I}_{\text{ref}}$ and later chunks progressively incorporate available observations without accessing future frames. During training, we sample global and causal conditions with equal probability, exposing the model to both complete multi-view geometry and the partial geometry available during autoregressive inference. We further augment conditions by using both raw and filtered point clouds, with a voxel resolution of $R = 128$ and a connectivity radius of $r = 2$ for floater removal. Foreground meshes are generated with RecGen (Zadaianchuk et al., 2026), and 20% of training samples are augmented with GPT-6 Astra (OpenAI, 2026) agentic meshes.

A.4 Data Curation

We train on DL3DV-10K (Ling et al., 2024), retaining every third frame, which yields approximately 1,000-1,500 frames per video. Videos are divided into 81-frame chunks, with adjacent chunks sharing one boundary frame. Camera intrinsics and poses are estimated with ViPE (Huang et al., 2025a), and all reconstructed geometry is aligned within the resulting shared scene coordinate system. Point clouds are reconstructed with Pi3 (Wang et al., 2026c), and video captions are generated with GPT-6. The point clouds and meshes are rendered along the estimated camera trajectories to produce RGB-D conditions paired with the target frames. We processed full 8,965 scenes from DL3DV using RecGen and a subset of approximately 5,246 scenes using Astra. Astra failed to produce any object meshes for only 0.4% of the processed scenes, while the corresponding failure rate for RecGen was approximately 3.5%. Using both mesh-generation methods reduced the proportion of scenes without any meshes to less than 0.3%, which means that over 99.7% of scenes have at least one mesh available for conditioning.

Table 5: **Mesh Curation Analysis.**

Count per scene	RecGen	Astra
0	320	23 failures
1	407	0
2	443	4
3	510	6
4	458	5
5	521	18
6	491	25
7	466	28
8	413	48
9	402	61
10	391	87
11–15	1,673	802
16–20	1,187	1,233
21–30	1,007	1,800
31–50	252	811
51–100	24	214
101–200	0	81
201–429	0	23
Counted scenes	8,965	5,246
Mean count	11.39	26.94
Maximum count	83	429

A.5 Training Hyperparameters

We finetune the control adapter with the flow-matching velocity loss, using 1,000 training timesteps with a schedule shift of 5.0. We use AdamW with a learning rate of 1×10^{-5}, a linear warmup over the first 100 iterations followed by a constant learning rate, $(\beta_1, \beta_2) = (0.9, 0.999)$, a weight decay of 0.01, and gradient clipping at a norm of 1.0. To support classifier-free guidance, we drop the text prompt with probability 0.1. We also drop the visual controls with probability 0.05. Training runs on 64 NVIDIA A100 GPUs, for a global batch size of 64 for approximately 10000 iterations.

Humanoid Setup and Application Details. Following GRAIL Xie et al. (2026), we first generate a reference human–object interaction (HOI) video of sitting, reconstruct the corresponding 4D HOI trajectory, and retarget the human motion to Unitree G1. We then apply GRAIL's scene-aware RL tracking method to assess whether the re-

Table 6: **Robustness Analysis.**

Configuration	Conditioning PSNR ↑	PSNR ↑	SSIM ↑	LPIPS ↓
1 anchor	11.66	14.26	0.3617	0.5161
2 anchors	11.74	14.36	0.3625	**0.5122**
Full Mesh + Full PLY (3 anchors)	**11.81**	**14.41**	**0.3658**	0.5141
Half Mesh + Full PLY	11.69	14.33	0.3616	0.5231
Full Mesh + PLY (conf. ≥ 0.3)	9.490	14.13	0.3540	0.5229
Mesh Position Shift of 5% + Full PLY	11.71	14.18	0.3564	0.5310

Table 7: **Runtime and Memory Comparison.**

Method	Model Time (min) ↓	Total Time (min) ↓	Peak Memory (GiB) ↓
Ours	157.6	166.4	44.76
Gen3C	**38.9**	**41.0**	**21.98**
AnyRecon	50.6	54.1	40.38
VerseCrafter	148.1	172.6	54.87
Lyra2	208.7	225.9	66.73
NeoVerse	239.1	254.8	46.54
Voyager	457.1	514.3	42.28

Table 8: **Runtime breakdown.**

Component / Step	Time (min)	Description / Details
Video diffusion denoising	150.3	13 chunks, 30 denoising steps each
Other processing across chunks	7.31	Input preparation, text/VAE processing, video encoding and saving
Processing outside chunks	4.06	Model initialization, initial and cumulative Pi3 control updates, stitching and validation
First-frame Pi3	1.79	Including model loading
First-frame SAM3 segmentation	0.74	Including model loading
First-frame RecGen	1.93	Model loading and mesh generation
Voxel-based filtering	0.30	Point-cloud filtering
Total	**166.4**	**Approximately 2 hours 46 minutes**
Astra: 1 iteration (optional)	2.89	Excluded from the total
Astra: 2 iterations (optional)	6.23	Excluded from the total

targeted motion can be executed in the corresponding chair scene, fine-tuning a pretrained SONIC Luo et al. (2026) whole-body controller with PPO in Isaac Lab. The policy conditions on reference motions, proprioceptive and previous-action histories, chair-relative state, and a CNN-encoded local height map. The controller and scene encoder are fine-tuned together, with the chair represented as fixed collision geometry. Training rewards encourage reference tracking of root/body positions, orientations and velocities, together with five-point positional tracking at the torso, wrists and ankles. Smoothness, joint-limit and undesired-contact penalties regularize the behavior, while an auxiliary motion-reconstruction loss stabilizes training and early termination handles excessive tracking error. In Fig. 8, geometric parameters include object dimensions, surface curvature, component thickness, and structural details, such as wheel radii or chair-leg shapes. Spatial parameters control position, orientation, and scale, while appearance parameters include color, roughness, and metallic properties. An initial mesh provides geometric guidance for VivaWorld's video generation. The generated views provide additional visual guidance about silhouettes, object parts, and spatial relationships. Astra uses these views as visual references to revise the procedural mesh-generation code, refining proportions, surface details, component placement, and materials. The revised mesh is then rendered from multiple viewpoints to inspect its consistency before guiding another round of video generation.

B EXPERIMENTS

Camera Evaluation For long-video style score, we compare consecutive anchors at per chunk's final frame, using the same anchors for every method. This measures local appearance stability and may miss gradual global drift, which differs from comparing only the first and last frames. Single-chunk and first-chunk style retain their endpoint comparisons within the evaluated windows. The

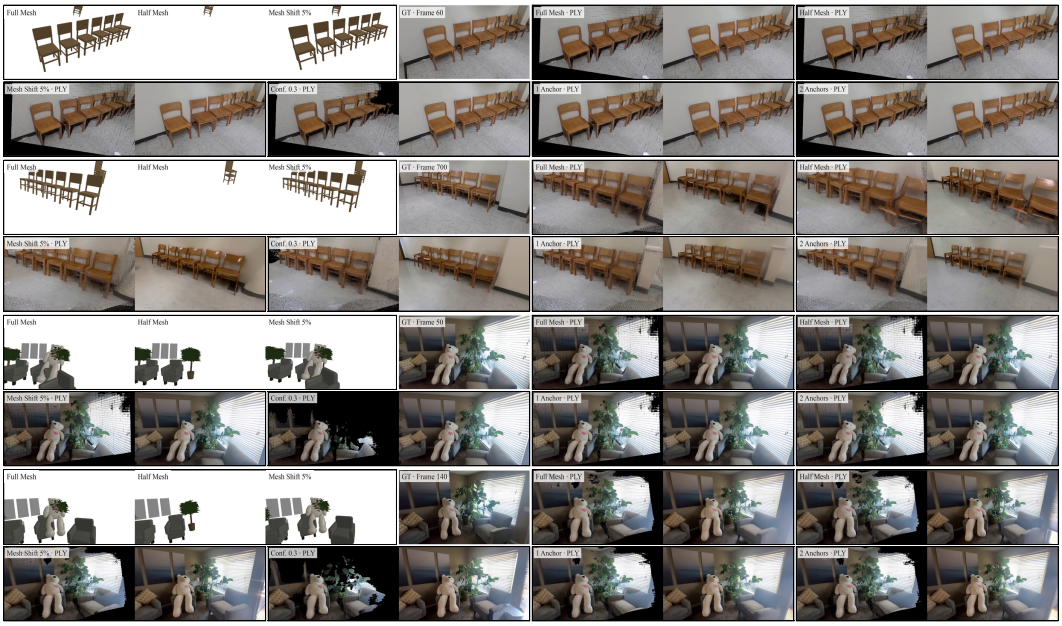

Figure 10: **Robustness.** We compare full conditioning (including imperfect meshes, such as extra chairs or a misplaced bear) with missing meshes, mesh misalignment, incomplete point clouds, and fewer object anchors (left: point-cloud rendering; right: predicted frame). All variants generate videos with good overall visual quality. Accurate point-cloud guidance prevents incorrect background chair meshes from causing artifacts in the final video. Insufficient mesh guidance can distort fine structures such as chair legs, while incomplete point clouds reduce foreground and background detail. Fewer anchors can work well, as training includes scenes with only one or two anchors, but may cause object drift in multi-object scenes.

reference trajectory is the ViPE trajectory of the ground-truth video. We use ViPE rather than the dataset-provided poses because it provides dense per-frame poses and intrinsics with one estimator for both DL3DV and Tanks and Temples. The trajectory of each generated window is estimated with DROID-SLAM, the camera pose estimator of WorldScore, using the reference per-frame intrinsics, so only extrinsics are estimated. Both trajectories are expressed relative to the first camera of the window. We first normalize the reference translations by the maximum displacement of the reference camera from its first position within the window, so that translation errors are a fraction of the trajectory extent rather than a metric distance. Rotations are compared without further alignment, while the monocular scale ambiguity of the estimated trajectory is removed by a single positive global scale factor, chosen to minimize the sum of per-frame Euclidean distances between the scaled estimated and the normalized reference camera positions. In addition, Cam. applies clipping and nonlinear normalization to each scene's errors before averaging: rotation errors above 15° and translation errors above 0.5 are clipped to these thresholds, so further increases do not affect the score. RotErr is the geodesic angle between the estimated and reference camera rotations, measured in degrees, and TransErr is the Euclidean distance between the scaled estimated and the normalized reference camera position.

Upstream Error Analysis. As shown in Table 6, we construct failure cases on 5 randomly selected test scenes by retaining approximately half the meshes (reducing the average from 9 to 4 objects per scene, the mesh number in 5 cases ranges from 3 to 17), filtering out 44.7% of the point-cloud (PLY) points with a confidence threshold of 0.3, and reducing the number of object anchors to 3, 2 and 1 to evaluate sensitivity to degraded upstream conditioning. Removing 44.7% point-cloud points substantially reduces conditioning PSNR (around - 2.2) and slightly decreases generation PSNR (-0.20) , while shifting mesh positions with 5% offset degrades performance a little as well. Using one or two anchors causes relatively small changes (since in our training dataset, some scenes only

have 1-2 anchors), and retaining half the meshes has a modest impact. This tolerance to incomplete geometry is supported by our training augmentations and the varying mesh coverage provided by sparser RecGen meshes (might introduce some inaccurate mesh as well for the feed forward curation in the training dataset) and denser Astra meshes (might introduce some coarse mesh shape), using progressive point-cloud reconstruction from generative frames (might have inaccurate point cloud conditions) and full ground truth global point clouds.

Runtime and Memory. On a single NVIDIA A100 80GB GPU, our method generates 1,000 frames at 832×480 resolution in 166.4 minutes, with a peak memory usage of 44.76 GiB. At the same resolution, our method is faster than VerseCrafter, Lyra2, NeoVerse, and Voyager, but slower than Gen3C and AnyRecon. Diffusion denoising accounts for 150.27 minutes, other processing across chunks takes 7.31 minutes, and preprocessing and remaining overhead take 8.82 minutes. Tables 7 and 8 provide the runtime comparison and detailed breakdown, respectively.

Table 9: **Evaluation on 3D Scene Reconstruction with DAv3 and 3DGS optimization.**

	Short-Video				Long-Video			
	SSIM ↑		LPIPS ↓		SSIM ↑		LPIPS ↓	
Method	Before	After	Before	After	Before	After	Before	After
AnyRecon	0.461	0.496	0.449	0.462	0.444	0.431	0.673	0.633
Gen3C	0.472	0.544	0.448	0.446	0.418	0.394	0.676	0.637
Voyager	0.461	0.504	0.492	0.492	0.350	0.276	0.802	0.828
Lyra2	0.472	0.555	0.439	0.426	0.423	0.406	0.658	0.616
VerseCrafter	0.451	0.495	0.475	0.446	0.403	0.383	0.763	0.748
NeoVerse	0.470	0.462	0.460	0.479	0.415	0.330	0.781	0.738
Ours	**0.480**	**0.569**	**0.417**	**0.406**	**0.447**	**0.435**	**0.598**	**0.568**

C More Comparison Results and Showcases

C.1 Comparison

3D Scene Evaluation Following Lyra2, we use Gaussians predicted by DA3 (Depth Anything 3). We additionally evaluate performance after 3DGS optimization, with results reported in Table 9. 3DGS optimization generally improves short-video SSIM but reduces long-video SSIM, possibly because accumulated inconsistencies make long sequences harder to fit with a single static scene. Our method retains the best SSIM and LPIPS in both settings, before and after optimization.

Qualitative Comparison. We provide additional qualitative comparisons of long video generation and 3DGS scene reconstruction in Fig. 11 and Fig. 12, following the same setting as in Sec. 4.3 (Qualitative Eval). Compared with the baselines, our method better preserves the object layout, appearance, and camera trajectory throughout long video generation. When the generated videos are used for 3DGS scene reconstruction, our method also yields cleaner and more coherent geometry, whereas the baselines suffer from blur, duplicated structures, and floating artifacts.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	Cam. ↑	Style ↑
Ours	**13.69**	**0.3797**	**0.5452**	**70.24**	**97.41**
Lyra2	13.17	0.3770	0.6534	58.46	89.60

Comparison on Landscape and Autonomous Driving (AV) Scenes To further demonstrate our method's performance with unreliable mesh conditioning, we select five natural scenes from DL3DV, including lakes and mountains, where we only have coarse meshes with incorrect shapes or placements. We use these imperfect Astra meshes during inference. We also evaluate five more challenging autonomous-driving scenes from OpenDV Yang et al. (2024), which is outside the DL3DV dataset used to train both our method and Lyra2 and we have no mesh rendering in some cases. For these driving scenes, we use Astra to update the mesh conditioning per chunk as the camera moves forward. We compare with the strongest baseline, Lyra2, on all ten scenes. The right table shows that our method achieves better image quality, camera control,

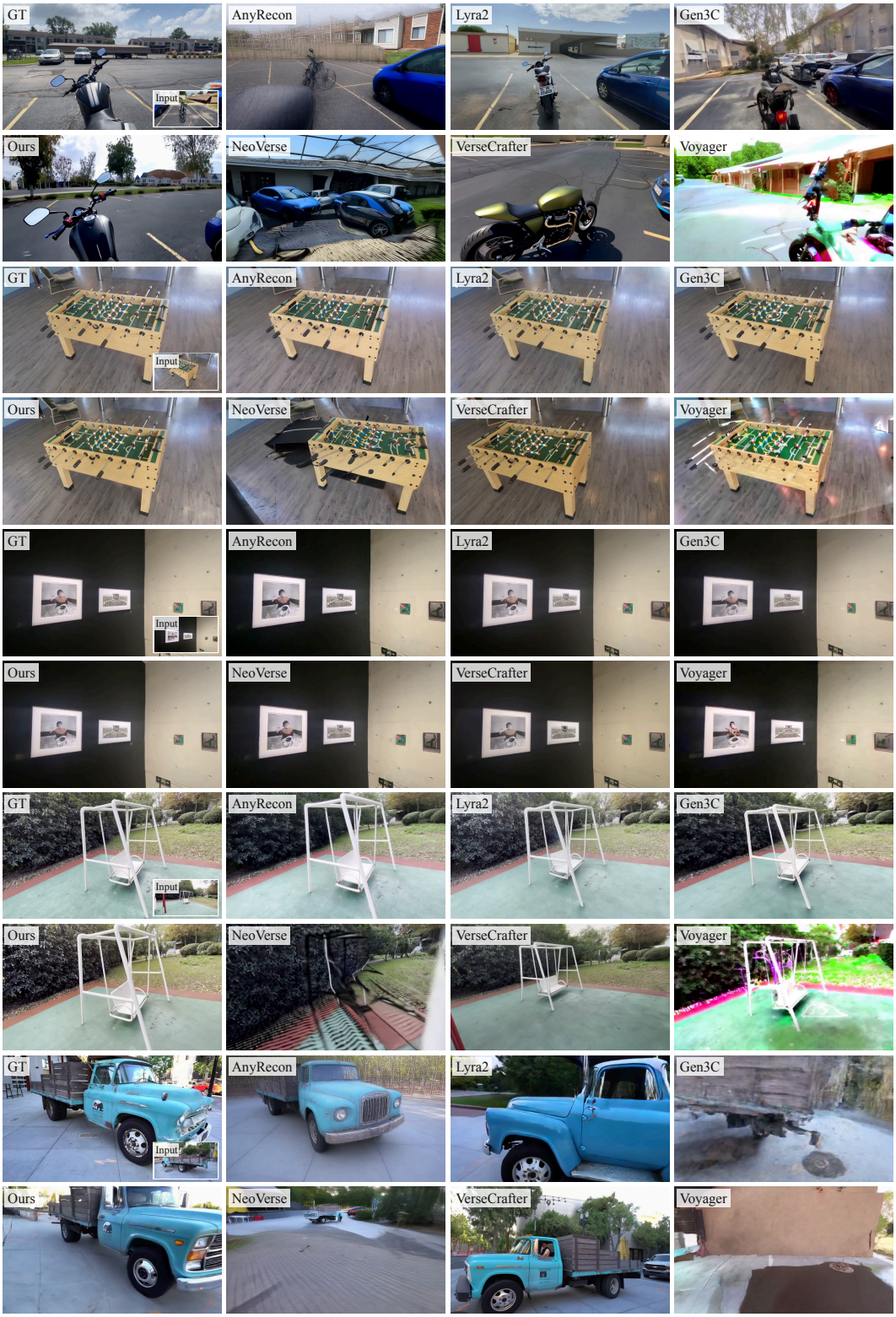

Figure 11: **Additional Video Generation Results.**



Figure 12: **Additional 3D Scene Reconstruction Results.**

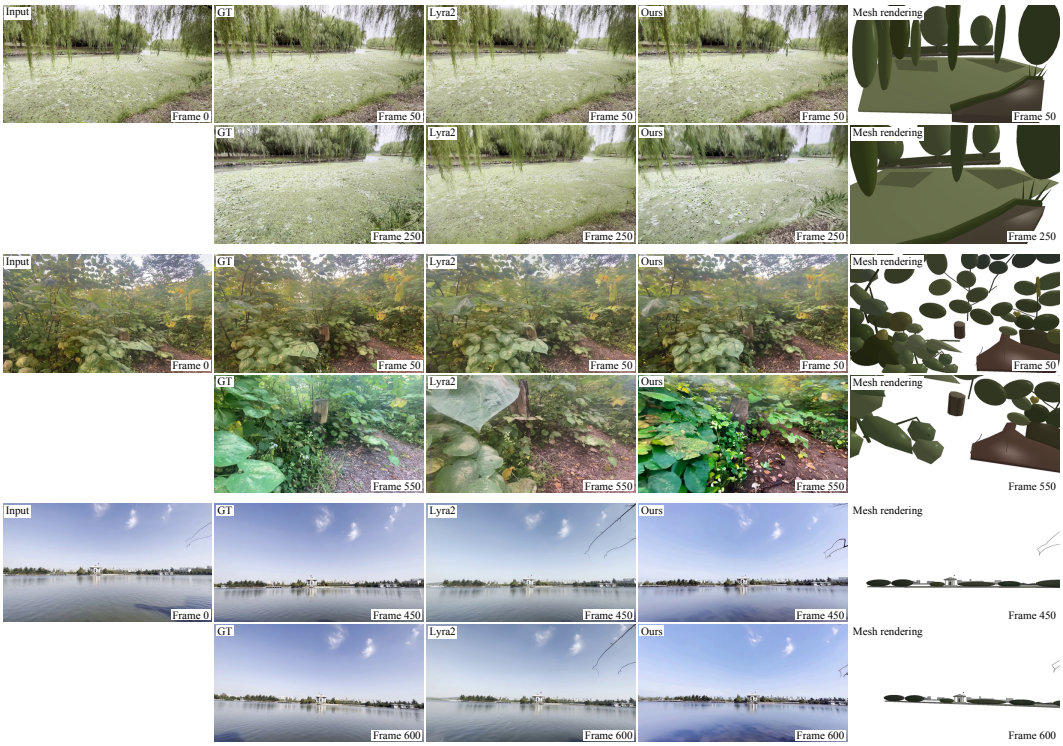

Figure 13: **Additional Video Generation on Landscape Scenes.** From left to right: input, GT, Lyra2, Ours, and mesh conditioning. Mesh panels show the corresponding first-frame mesh guidance.

and style consistency, even with unreliable meshes featuring coarse shapes or incorrect placements. We provide selected additional qualitative comparisons on landscape and driving scenes in Figs. 13 and 14. The landscape examples use fixed initial Astra meshes and the driving examples show prediction-derived per-chunk Astra mesh updates.

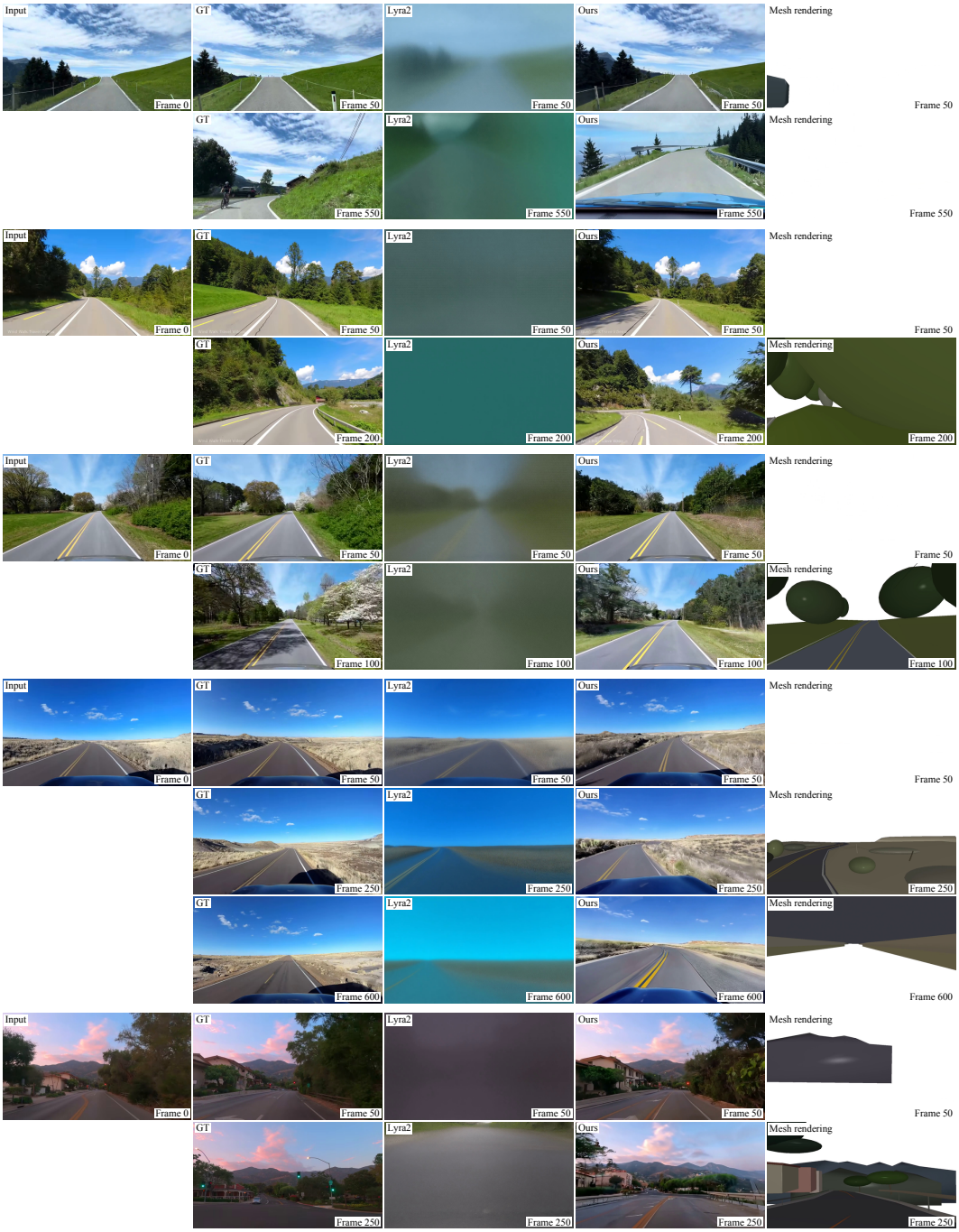

Figure 14: **Additional Video Generation on Driving Scenes.** From left to right: input, GT, Lyra2, Ours, and prediction-derived mesh conditioning. Mesh panels show the corresponding per-chunk Astra guidance.

C.2 More Showcases

We provide additional results on long video generation and 3DGS reconstruction following the same setting in the qualitative study. Our method achieves accurate camera control and strong geometric consistency in long video generation. Moreover, when the generated long videos are used for 3D reconstruction, our method still maintains strong consistency in geometry, shape, and camera trajectory.



Figure 15: **Additional Video Generation and 3D Scenes.** For each case: input image and ground-truth frames (top), our generated video (middle), and our 3DGS reconstruction rendered along the same trajectory (bottom). Numbers denote seconds.



Figure 16: **Additional Video Generation and 3D Scenes.** For each case: input image and ground-truth frames (top), our generated video (middle), and our 3DGS reconstruction rendered along the same trajectory (bottom). Numbers denote seconds.

To further demonstrate robustness to unusable meshes, we show two DL3DV scenes where RecGen fails, using only Astra meshes with incorrect shapes and placements (top two cases) and four cases using only progressively reconstructed point clouds and global camera conditioning. In Fig. 17, our model still generates realistic videos, although quality may be lower than with the full pipeline.

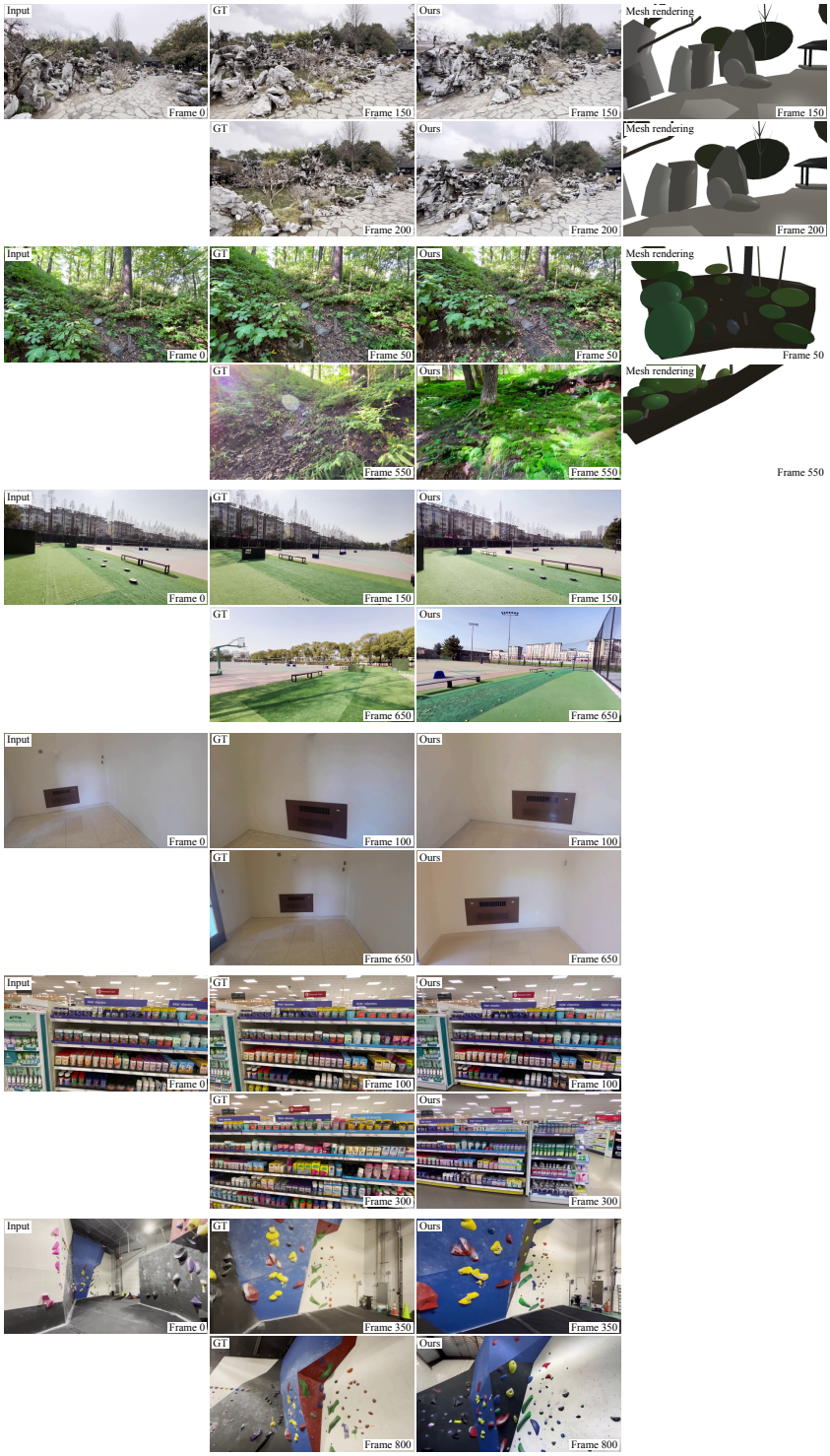

Figure 17: **Additional Video Generation.** From left to right: input, GT, Ours, and mesh conditioning. Blank mesh cells in the last four scenes indicate only point-cloud guidance.