Lyra 2.0: Explorable Generative 3D Worlds

Figure 1: Lyra 2.0 enables long-horizon 3D-consistent scene generation from a single image. Starting from an input image, users iteratively define camera motion to explore the scene, while Lyra 2.0 synthesizes spatially persistent video outputs that progressively expand the environment. These videos can be directly reconstructed into high-fidelity 3D Gaussians and surface meshes, yielding 3D assets deployable in simulation engines and interactive viewers.

Abstract

Recent advances in video generation enable a new paradigm for 3D scene creation: generating camera-controlled videos that simulate scene walkthroughs, then lifting them to 3D via feed-forward reconstruction techniques. This generative reconstruction approach combines the visual fidelity and creative capacity of video models with 3D outputs ready for real-time rendering and simulation. Scaling to large, complex environments requires 3D-consistent video generation over long camera trajectories with large viewpoint changes and location revisits, a setting where current video models degrade quickly. Existing methods for long-horizon generation are fundamentally limited by two forms of degradation: spatial forgetting and temporal drifting. As exploration proceeds, previously observed regions fall outside the model's temporal context, forcing the model to hallucinate structures when revisited. Meanwhile, autoregressive generation accumulates small synthesis errors over time, gradually distorting scene appearance and geometry. We present Lyra 2.0, a framework for generating persistent, explorable 3D worlds at scale. To address spatial forgetting, we maintain per-frame 3D geometry and use it solely for information routing—retrieving relevant past frames and establishing dense correspondences with the target viewpoints—while relying on the generative prior for appearance synthesis. To address temporal drifting, we train with self-augmented histories that expose the model to its own degraded outputs, teaching it to correct drift rather than propagate it. Together, these enable substantially longer and 3D-consistent video trajectories, which we leverage to fine-tune feed-forward reconstruction models that reliably recover high-quality 3D scenes.

1. Introduction

Trained on massive internet data, video diffusion models [6, 13, 90, 107] now exhibit remarkable visual fidelity and strong local 3D consistency between neighboring frames. This progress enables generative reconstruction [2]: given a single image and a prescribed camera trajectory, a video diffusion model synthesizes dense novel views that serve as virtual captures for feed-forward 3D reconstruction, recovering explicit scene geometry and appearance. By replacing labor-intensive real-world capture with generative view synthesis, this paradigm enables scalable creation of diverse, high-quality, and even entirely imaginary 3D environments.

However, scaling generative reconstruction to large, complex environments—such as navigating across rooms or long city streets—requires maintaining 3D consistency over extended trajectories with substantial viewpoint changes and revisits. Current video models generate frames autoregressively and struggle in such unbounded exploration scenarios, primarily suffering from two forms of degradation. First, spatial forgetting: as the camera moves, previously observed regions inevitably exceed the model's finite temporal context window. Upon revisiting these areas, the model is forced to hallucinate structures from scratch, breaking global layout consistency. Second, temporal drifting: autoregressive generation is inherently susceptible to error accumulation. Small per-step synthesis artifacts compound over time, leading to severe color shifts and structural distortions. This is further exacerbated during camera exploration, where continuously introduced unseen regions diminish visual overlap with early history frames, depriving the model of reliable geometric and texture constraints.

Recent efforts to mitigate spatial forgetting incorporate historical memory into the generation process. A prominent line of work [2, 81, 139, 141] maintains a cumulative 3D representation, conditioning subsequent frames on rendered views of the reconstructed geometry. While providing explicit spatial constraints, this tightly coupled design suffers from error amplification: generative artifacts degrade the 3D geometry, which in turn produces flawed conditioning for future frames. Alternatively, incorporating history frames directly into the context window via camera pose embeddings [26] avoids corrupted 3D intermediaries. Yet, this relies entirely on the model's self-attention to infer long-range geometric correspondences, which frequently fails under large viewpoint variations. Instead, we bridge these two memory mechanisms by decoupling

geometric tracking from pixel synthesis. We utilize an explicit 3D proxy solely for information routing—retrieving relevant historical context and establishing spatial correspondences. Given this undistorted history context with dense spatial grounding, the actual novel view synthesis is left to the diffusion model's learned pixel prior, which resolves geometric inconsistencies and synthesizes novel views without propagating hard rendering artifacts.

To mitigate temporal drifting, existing strategies typically extend the temporal context length to anchor on past frames [135]. However, in scene exploration, camera motion inherently moves early frames out of the field of view, rendering long-context anchoring ineffective for suppressing drift in newly observed areas. We propose alleviating the underlying training-inference discrepancy through a self-augmentation training scheme. By stochastically conditioning the network on its own one-step denoised predictions during training rather than perfect ground-truth frames, we expose the model to the exact error distributions encountered during autoregressive inference. Together with retrieving high-overlap history frames in the context window, the video model learns to actively mitigate drifting in recent generations with minimal computational overhead.

Equipped with these mechanisms, our model achieves highly persistent and long-horizon scene generation. Nevertheless, videos synthesized by diffusion models inevitably contain minor multi-view inconsistencies that easily break traditional 3D reconstruction models, causing floaters and noisy artifacts. To achieve reliable scene reconstruction, we employ a feed-forward 3D Gaussian Splatting (3DGS) pipeline [58]. Fine-tuned on our generated sequences, this feed-forward model leverages its learned multi-view prior to tolerate minor inconsistencies, effectively bridging the domain gap and producing clean, coherent 3D structures.

We integrate these components into Lyra 2.0, an interactive system for large-scale 3D scene exploration. Starting from a single image, Lyra 2.0 empowers users to define arbitrary long-horizon camera trajectories and progressively reconstruct complex environments. As demonstrated in Fig. 1, our approach supports extensive scene navigation, including lookbacks and large-scale synthesis. The generated content can then be reliably reconstructed into high-quality 3D Gaussians and surface meshes with accurate geometry, ready for downstream applications in embodied AI and immersive rendering.

2. Related Work

Camera-Conditioned Video Generation. There have been significant advances in extending video diffusion models to incorporate camera control. Early approaches inject explicit camera parameterizations into the generative backbone. For instance, MotionCtrl [114] flattens per-frame camera pose matrices into vectors and injects them into intermediate feature representations of a pre-trained video diffusion model. Subsequent works [3, 4, 26, 120] adopt dense ray-based encodings using Plücker coordinates [10, 93], enabling pixel-wise camera conditioning and improved viewpoint control. Following the success of Genie 3 [5], an increasingly popular line of work [28, 50, 70, 100, 138] formulates camera control as an action-conditioning problem, where viewpoint changes are driven by discrete control signals such as keyboard inputs. To further improve geometric faithfulness, recent approaches [48, 81, 117, 129, 130, 139] introduce more structured 3D guidance signals beyond per-frame pose conditioning. These methods condition generation on renderings of estimated 3D geometry, such as global point cloud renderings or depth-warped images, to better constrain spatial structure during generation. GenWarp [88] introduces correspondence-based conditioning but is limited to single-image diffusion models.

While these works produce compelling videos under viewpoint control, the underlying 3D consistency of the generated scenes is often limited and does not remain persistent when revisiting previously generated regions. Our work builds upon the camera-controlled video generation paradigm and addresses the fundamental problems of spatial forgetting and temporal drifting in long-horizon 3D-consistent generation.

Memory-Aware Long Video Generation. Although camera conditioning enables controllable viewpoint changes,

most video diffusion models remain constrained by a fixed temporal context window. As a result, long-horizon consistency degrades once previously generated content falls outside the attention span of the model. To address this limitation, recent works augment generative models with explicit memory mechanisms. A first family of approaches [24, 51, 118, 128] relies on retrieval-based memory. These methods treat past frames as an external memory bank and dynamically select relevant observations to guide the next generation. For example, Context-as-Memory [128] and WorldMem [118] retrieve earlier frames based on field-of-view (FOV) overlap, while VMem [51] performs geometry-aware retrieval using indexed 3D surface elements instead of purely view-based similarity. A second line of work [48, 62, 116, 139, 141] enforces spatial persistence through explicit 3D representations accumulated over time. Rather than retrieving individual frames, these methods construct and maintain a global scene structure that serves as a unified memory for camera control and revisit consistency. A third direction [15, 32, 33, 74, 84, 137] improves long-range temporal coherence by modifying the internal architecture of the generator, maintaining persistent latent states or key–value caches that propagate information across timesteps. Orthogonally, FramePack [135] compresses history frames into compact contextual slots through variable patchification based on temporal relevance, extending the effective context window without architectural changes.

In contrast to global 3D memory methods that rely on a single accumulated scene representation, we maintain per-frame 3D geometry and use it exclusively for information routing, i.e., retrieving relevant history frames and establishing dense geometric correspondences, while relying on the video model's generative prior for appearance synthesis. Combined with a self-augmentation strategy that mitigates temporal drifting, our approach enables scalable scene expansion and long-horizon 3D consistency under complex camera motion.

3D Scene Generation. A prominent line of work [9, 58, 69, 80, 95, 134] reconstructs 3D Gaussians [43] from one or multiple input views in a fully feed-forward manner. Recent approaches combine generative modeling with feed-forward 3D reconstruction to reduce the reliance on densely sampled multi-view inputs. Bolt3D [98], for example, trains a pointmap [111] autoencoder to generate multi-view pointmaps, which are subsequently used for feed-forward 3D reconstruction. Wonderland [54] utilizes a camera-controlled video diffusion model to synthesize multi-view imagery and then predicts 3D Gaussians with a dedicated feed-forward network. More recently, Lyra [2] adopts a camera-controlled video model as a teacher within a self-distillation framework, enabling the training of a student 3D reconstruction model without requiring real-world multi-view supervision. FlashWorld [52] further demonstrates efficient 3D scene generation using a distilled camera-controlled video diffusion model. WorldExplorer [85] generates navigable 3D scenes from text by iteratively producing camera-guided videos from panoramic initializations and fusing them into 3D Gaussians via per-scene optimization. Concurrently, Video-to-World [31] proposes a non-rigid alignment procedure to correct 3D inconsistencies in video generations before lifting into 3D. Free-Range Gaussians [89] tackles generative 3D reconstruction by using flow matching directly on the Gaussian parameters.

These methods achieve strong results but typically remain limited in view coverage. We instead generate long, 3D-consistent videos from a single image and lift them into large-scale 3D Gaussians and meshes via a scalable feed-forward reconstruction pipeline.

3. Preliminaries

DiT-Based Latent Video Diffusion. Our method builds upon DiT-based latent video diffusion models [13, 107]. Given an RGB video of F frames, $\mathbf{x} \in \mathbb{R}^{F\times 3\times H\times W}$, a VAE encoder compresses it into a latent $\mathbf{z} = \mathcal{E}(\mathbf{x}) \in \mathbb{R}^{F'\times C\times h\times w}$, from which a decoder reconstructs $\hat{\mathbf{x}} = \mathcal{D}(\mathbf{z})$. To jointly handle images and videos, modern causal video VAEs encode the first frame independently and temporally compress subsequent frames. We adopt the Wan 2.1 VAE [107], which downsamples 8× spatially and 4× temporally, giving $F' = \lfloor (F-1)/4 \rfloor + 1$, $C = 16$, $h = H/8$, $w = W/8$. Generation is performed in this latent space via flow matching [61]: given a clean latent $\mathbf{z}_0$ and noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$, we form $\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}$ for $t \in [0, 1]$ and train a DiT v_θ to regress

the velocity:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\, t,\, \boldsymbol{\epsilon}} \left[\left\| v_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{z}_0) \right\|^2 \right], \tag{1}$$

where $\mathbf{c}$ denotes conditioning signals (e.g., text). Long videos can be produced by generating fixed-length segments autoregressively, conditioning each step on previously generated frames.

Camera-Conditioned Video Generation. Generating 3D-consistent scene explorations requires the model to follow a prescribed camera trajectory. For the i-th image, we denote the world-to-camera extrinsic as $\mathbf{T}_i = [\mathbf{R}_i \mid \mathbf{t}_i] \in \mathbb{R}^{3\times4}$, intrinsic $\mathbf{K}_i \in \mathbb{R}^{3\times3}$, and estimated depth map $D_i \in \mathbb{R}^{H\times W}$ [81]. Two complementary strategies exist for injecting camera information into a DiT. Depth-based warping [81] forward-warps the most recent frame I_j to each target viewpoint $(\mathbf{T}_i, \mathbf{K}_i)$ using its depth D_j, encodes and concatenates the renderings with the denoising latent. We find that within Wan 2.1 [107], this mechanism alone already delivers accurate camera control even along long trajectories. However, when the viewpoint change is large enough that no warped pixels land on the target view, the control signal is lost entirely, and the visual quality degrades significantly. We therefore complement it with Plücker ray injection [26], which computes 6D ray coordinates $\mathbf{r}_i(u, v) = (\mathbf{d},\ \mathbf{o} \times \mathbf{d}) \in \mathbb{R}^6$ per pixel, projects them to the DiT's hidden dimension via an MLP, and adds them to token features, providing an extra hint in case of drastic viewpoint changes.

Context Compression via FramePack. We adopt FramePack [135] to compress the history context and mitigate drifting. We describe its details here and discuss additional strategies to further reduce drifting in § 4.3. FramePack compresses history frames with variable patchification kernels by temporal proximity: recent frames use a small kernel for fine-grained tokenization, while distant frames use a large kernel for aggressive compression. This allows the model to attend to a long temporal horizon within a fixed token budget. Typically, the temporal history is organized as:

$$\underbrace{\texttt{f1k1}}_{\text{anchor}}\ \underbrace{\texttt{f16k4}\ \ \texttt{f2k2}\ \ \texttt{f1k1}}_{\text{temporal slots}}\ \ \underbrace{\texttt{g20}}_{\text{generate}}, \tag{2}$$

where $\texttt{f}n\texttt{k}m$ denotes n frames compressed with spatial subsampling factor m, and $\texttt{g20}$ is the 20-frame generation target. The anchor frame (the initial image I_0) is always included at full resolution as a fixed reference point, serving as an early-established endpoint [135] that prevents the model from drifting away from the original scene appearance.

4. Method

4.1. Overview

Given a single input image I_0 and a camera trajectory $\{(\mathbf{T}_i, \mathbf{K}_i)\}_{i=0}^{T-1}$, our goal is to generate a long, camera-controlled video that maintains global 3D consistency across all frames and can be lifted into an explorable 3D scene.

As illustrated in Fig. 2, Lyra 2.0 generates long videos through an autoregressive retrieve–generate–update loop. At each iteration, the user first provides a 3D camera trajectory and an optional text prompt to guide outpainting. Then we (i) retrieve history frames whose 3D content is most relevant to the target viewpoint, (ii) generate the next video segment conditioned on both temporal history and retrieved spatial context, and (iii) update the memory with the newly generated frames. At the core of this pipeline are two mechanisms that address the key challenges in long-horizon autoregressive generation: anti-forgetting (§ 4.2), which builds a spatial memory with per-frame 3D geometry and bridges it with the video model's context to maintain spatial consistency when revisiting previously explored regions, and anti-drifting (§ 4.3), which adaptively compresses history frames and mitigates quality degradation over long sequences. The memory grows with each iteration step, enabling the model to maintain consistency over arbitrarily long trajectories and across

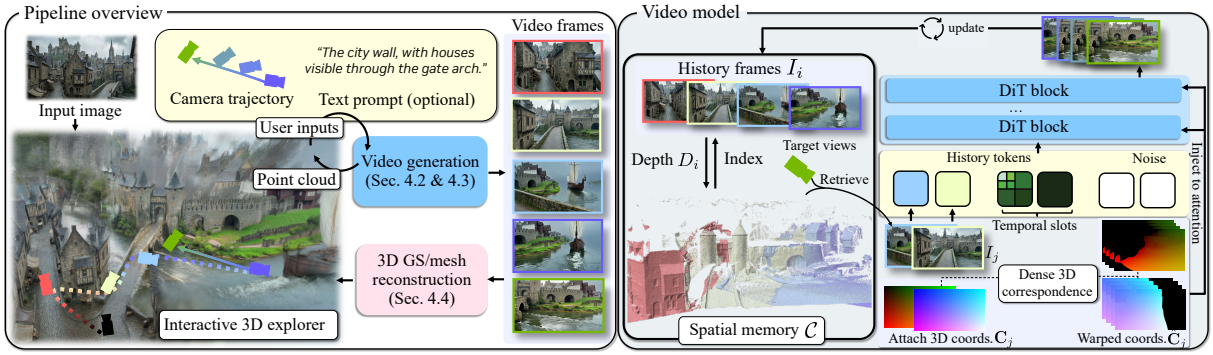

Figure 2: Method overview. (Left) Given an input image, Lyra 2.0 iteratively generates video segments guided by a user-defined camera trajectory from an interactive 3D explorer and an optional text prompt, lifting each segment into 3D point clouds fed back for continued navigation. Generated video frames are finally reconstructed and exported as 3D Gaussians or meshes. (Right) At each step, history frames with maximal visibility of the target views are retrieved from the spatial memory. Their canonical coordinates are warped to establish dense 3D correspondences and injected into DiT via attention, together with compressed temporal history.

revisits to previously explored regions. The generated long video is then lifted into explicit 3D representations via feed-forward 3D reconstruction (§ 4.4).

4.2. Anti-Forgetting for 3D-Persistent Video Generation

Achieving long-range spatial consistency requires recalling geometrically relevant history observations regardless of their temporal distance. Our core intuition is to use noisy 3D geometry estimation exclusively for information routing—selecting which history observations are relevant and establishing geometric correspondence between history and future viewpoints—while the video model handles all appearance synthesis and resolves inconsistencies between observations. Following this intuition, we first build a 3D cache that stores per-frame geometry information, and design a retrieval strategy that selects the most informative history frames for a given target viewpoint to condition the video model.

Building the 3D Cache. We maintain a 3D cache $\mathcal{C}$ that grows incrementally as the video is generated. For each frame I_i with estimated depth D_i [58] and camera intrinsic and extrinsic $(\mathbf{T}_i, \mathbf{K}_i)$, our 3D cache maintains two components: (i) the full-resolution depth map D_i and camera parameters; (ii) a downsampled point cloud $\mathbf{P}_i \in \mathbb{R}^{(H/d)\times(W/d)\times 3}$, obtained by subsampling the depth map by factor d and unprojecting it into world coordinates. This first component preserves full geometric precision for correspondence computation, while the second one is exclusively used for efficient retrieval.

Critically, the cache stores the geometry of each frame independently, and we never fuse them into a single global point cloud. This is particularly important in long-horizon generation, where depth estimation quality inevitably degrades over time, since it runs on the generated frames rather than real images. By maintaining per-frame point clouds, we avoid accumulating cross-view misalignment from imperfect depth into a single corrupted reconstruction.

Geometry-Aware Retrieval. Since the context window of a video model is limited, selecting the most informative history frames is critical for maximizing long-range consistency and efficiency. At each autoregressive step, we select N_s history frames whose 3D content is most visible from the target viewpoint. To achieve this, we compute the visibility score ϕ of each history frame. Specifically, given a target camera $(\mathbf{T}^*, \mathbf{K}^*)$, we project every downsampled point cloud $\mathbf{P}_i$ onto the target image plane. Then, for each pixel on the

target image plane, we compute the minimum projected depth over all frames to handle occlusion. A point is considered visible if and only if the difference between its depth and the minimum depth is less than a threshold δ. The visibility score $\phi(i)$ of frame i is the count of its visible points. During training, we sample the history frames proportional to visibility scores $\phi(i)$ to make the model robust to different frame retrieval results. At inference, we greedily maximize coverage: iteratively selecting the frame that covers the most not-yet-covered target pixels, up to N_s frames. This avoids redundant selection of nearby viewpoints and maximizes the collective spatial coverage.

With this mechanism, even when the camera revisits a region hundreds of frames later—far beyond the model's temporal context window—our retrieval can naturally recall the relevant observations via their 3D overlap.

Injecting Spatial Memory into the Video Model. Having retrieved the most relevant history frames $\{I_j\}_{j=0}^{N_s-1}$, we inject them into the video model as spatial slots: each retrieved frame is encoded independently by the VAE as image tokens (without temporal compression) and placed alongside the temporal FramePack slots and generation tokens. We apply the same variable-kernel spatial compression from FramePack to both the temporal and spatial slots. The full context layout is:

$$\underbrace{\texttt{f1k1}}_{\text{anchor}}\ \underbrace{\texttt{f4k2 f1k1}}_{\text{spatial slots}}\ \underbrace{\texttt{f16k4 f2k2 f1k1}}_{\text{temporal slots}}\ \underbrace{\texttt{g20}}_{\text{generate}},$$

where spatial slots contribute N_s=5 retrieved frames: 4 frames at subsampling factor 2 and 1 frame at full resolution. All tokens are jointly processed by the full DiT self-attention.

While prior retrieval-based approaches [118, 128] inject history frames in a similar fashion, they lack geometric grounding for precise multi-view alignment. To address this, we further establish dense correspondences via canonical coordinate warping: for the j-th retrieved frame, we assign a canonical coordinate map $\mathbf{C}_j \in [-1,1]^{3\times H\times W}$ whose three channels are $(u, v, 2\cdot\frac{j}{N_s}-1)$ where (u,v) encodes the normalized spatial position. We then forward-warp $\mathbf{C}_j$ using the full-resolution depth:

$$\hat{\mathbf{C}}_j = \texttt{FwdWarp}(\mathbf{C}_j,\ D_{s_j},\ \mathbf{T}_{s_j},\ \mathbf{T}^*,\ \mathbf{K}_{s_j},\ \mathbf{K}^*). \tag{3}$$

We additionally warp the depth as a fourth channel, yielding a 4-channel map $[\hat{\mathbf{C}}_j; \hat{D}_j]$ per retrieved frame. When fewer than N_s frames are retrieved, missing slots are padded so the model can distinguish real correspondences from empty slots. To feed the warped correspondence maps into DiT, we encode them via positional encoding and aggregate through a learned MLP. The output embeddings are added to the tokens at the self-attention layer of every transformer block.

Note that we warp canonical coordinates rather than RGB images for a specific reason: warped RGB inevitably contains disocclusion holes, stretching artifacts, and depth-boundary bleeding. If conditioned on such images, the video model tends to re-generate these artifacts—the warped image acts as a crutch that bypasses the generative prior rather than informing it. Canonical coordinates carry the same geometric correspondence information without any appearance content, leaving appearance synthesis entirely to the video model.

In summary, our video model context comprises three complementary signals: (1) the retrieved history frames $\{I_j\}_{j=0}^{N_s-1}$ encoded as spatial slots; (2) the forward-warped correspondence maps $[\hat{\mathbf{C}}_j; \hat{D}_j]_{j=0}^{N_s-1}$; and (3) the compressed temporal history via FramePack (Eq (2)).

4.3. Anti-Drifting for Long-Horizon Video Generation

The root cause of drifting is observation bias [135]: during training, the model conditions on ground-truth history frames, but at inference it must condition on its own imperfect outputs. This train-test discrepancy

means that per-step errors—color shifts, blurring, distortions—go uncorrected and compound across autoregressive steps, gradually degrading quality. Context compression via FramePack (§ 3) alleviates drift by extending the temporal horizon and anchoring generation to the original image, but it does not close the fundamental observation bias gap. We therefore complement it with a self-augmentation training strategy that directly reduces the train-test discrepancy.

Self-Augmentation Training. Related approaches such as Self-Forcing [36] mitigate drifting by conditioning the model on its own predictions during training, but are primarily designed for causal network architectures. Directly applying self-forcing to our bi-directional video model is prohibitively expensive: each history segment would require full bi-directional attention and multi-step denoising (e.g., 35 steps) to simulate the model's inference-time outputs.

To address this, we introduce a lightweight self-augmentation strategy. Consider an autoregressive training step with ground-truth history frames $\mathbf{x}^{\text{hist}}$ and current chunk frames $\mathbf{x}^{\text{cur}}$. Since our VAE is causal, encoding the current chunk depends on the temporal cache from the history segment. We encode both using clean ground-truth frames: $\mathbf{z}_0^{\text{hist}} = \mathcal{E}(\mathbf{x}^{\text{hist}})$ and $\mathbf{z}_0^{\text{cur}} = \mathcal{E}(\mathbf{x}^{\text{cur}} \mid \mathbf{x}^{\text{hist}})$, where the conditioning notation denotes the causal VAE cache dependency.

With probability p_{aug}, we corrupt the history latent by sampling $t \sim \mathcal{U}(0, 0.5)$ and adding noise according to the flow matching schedule:

$$\mathbf{z}_t^{\text{hist}} = (1-t)\,\mathbf{z}_0^{\text{hist}} + t\,\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{4}$$

The video model then performs one-step denoising to produce an approximate reconstruction:

$$\tilde{\mathbf{z}}_0^{\text{hist}} = \mathbf{z}_t^{\text{hist}} - t \cdot v_\theta(\mathbf{z}_t^{\text{hist}}, t, \mathbf{c}), \tag{5}$$

and we replace $\mathbf{z}_0^{\text{hist}}$ with $\tilde{\mathbf{z}}_0^{\text{hist}}$ as the DiT's conditioning context. Crucially, the target latent $\mathbf{z}_0^{\text{cur}}$ is always encoded with the clean history cache, and the flow matching loss supervises the DiT to denoise toward this clean $\mathbf{z}_0^{\text{cur}}$ despite receiving corrupted conditioning. This teaches the model to recover high-quality outputs from imperfect history context, effectively learning to counteract drifting artifacts during autoregressive inference. The overall overhead is minimal, requiring only one additional DiT forward pass.

4.4. 3D Reconstruction

We lift the generated videos from Lyra 2.0 into explicit 3D representations for downstream applications such as embodied AI simulation and virtual reality.

3D Gaussian Splatting. We adopt Depth Anything v3 (DAv3) [58], a feed-forward 3D foundation model that predicts per-pixel 3DGS attributes from input images. However, the pretrained DAv3 model exhibits two main limitations in our setting. First, DAv3 predicts one Gaussian per pixel, which leads to an excessively large number of Gaussians for the high-resolution inputs produced by our system. To address this, we modify the Gaussian DPT head in the DAv3 architecture to produce a feature map downsampled by a factor of $k \times k$. This allows the network to process the original high-resolution images while reducing the number of predicted Gaussians by k^2, yielding a more compact representation suitable for real-time rendering and data streaming. Second, DAv3 is not optimized for generated data, where small geometric inconsistencies are common. To improve robustness, we fine-tune the model on scenes generated by our video model. Similar to Lyra [2], this improves robustness to artifacts commonly present in generative data.

Surface Mesh Extraction. After obtaining the 3DGS, we further extract a surface mesh. Specifically, we develop a hierarchical sparse grid approach for large-scale mesh extraction based on OpenVDB [72, 115], allocating fine grid cells near the generation viewpoints and coarse cells in the distant background. The median depth from the Gaussian reconstruction is rasterized as a depth map in each view with normals

Table 1: Quantitative comparison on single-view to long video generation. Best results are shown in bold and second best are underlined.

Method	DL3DV							Tanks-and-Temples						
	SSIM↑	LPIPS↓	FID↓	Subjective Qual.↑	Style Consist.↑	Camera Ctrl.↑	Reproj. Err.↓	SSIM↑	LPIPS↓	FID↓	Subjective Qual.↑	Style Consist.↑	Camera Ctrl.↑	Reproj. Err.↓
GEN3C [81]	0.346	0.535	58.96	24.60	76.77	69.54	0.068	0.350	0.589	79.07	21.75	75.54	70.91	0.054
Yume1.5 [70]	0.342	0.719	84.84	22.80	66.73	–	0.095	0.348	0.702	89.69	28.68	78.63	–	0.083
CaM [128]	0.370	0.562	50.43	35.19	82.63	42.71	0.069	0.367	0.605	59.20	34.22	82.83	31.86	0.056
VMem [51]	0.331	0.744	120.59	18.54	76.14	0.68	0.268	0.338	0.767	136.48	16.21	70.54	0.00	0.263
SPMem [117]	0.383	0.522	53.77	38.32	82.79	62.05	0.074	0.383	0.571	60.11	34.41	79.68	45.07	0.059
HY-WorldPlay [37]	0.373	0.765	139.36	4.79	54.62	–	0.092	0.380	0.796	163.54	3.24	48.22	–	0.084
Ours	0.388	0.498	43.43	44.54	87.46	64.67	0.076	0.384	0.552	51.33	43.35	85.07	63.87	0.069
Ours DMD	0.359	0.507	43.63	45.21	88.57	65.64	0.088	0.362	0.545	49.71	43.02	78.91	58.12	0.077

computed as the gradient of depth, and we use the resulting oriented point cloud to construct a signed distance function on the sparse grid. Surfaces are extracted via marching cubes, stitched across hierarchy levels, and decimated for efficient downstream processing.

4.5. Distilled Model for Accelerated Inference

We additionally train a distilled version of our model using Distribution Matching Distillation (DMD) [126] to accelerate inference. Starting from our trained teacher model, we distill a student model that generates videos in 4 denoising steps instead of 35. We also distill the classifier-free guidance into the student, eliminating the need for separate conditional and unconditional forward passes at inference. During distillation, we retain our self-augmentation strategy so that the student remains robust to autoregressive error accumulation. Combined, the reduced sampling steps and single-pass inference reduce the per-step generation time by roughly 13× while maintaining comparable visual quality for interactive use cases.

5. Experiments

5.1. Training Details

Datasets. We train our model on DL3DV [60], which contains 10K long video clips of diverse real-world scenes. We estimate camera poses using ViPE [35] and predict per-frame depth with Depth Anything V3 [58]. Video captions are generated using Qwen3-VL-8B-Instruct [103].

Paired Data Curation. For real-world videos from DL3DV, we sample 1,000 frames per video. During training, we construct conditioning–target pairs using two complementary strategies. With 30% probability, we train in image-to-video (I2V) mode, where the model generates the first $L = 80$ consecutive frames conditioned on a single initial frame. With the remaining 70% probability, we perform autoregressive chunk-based training. Specifically, given a sequence of T frames, we uniformly sample a segment index $s \in [0, S_{\max})$, where $S_{\max} = \lfloor (T-1)/L \rfloor - 1$. The history window spans frames $[0,\ s \cdot L + 1)$ as conditioning context, and the ground-truth target consists of the next L consecutive frames in segment $s + 1$.

5.2. Evaluation on Long Video Generation

Baselines and Metrics. We compare against recent camera-controllable long video generation methods with memory mechanisms: Yume-1.5 [70], GEN3C [81], Context as Memory (CaM) [128], VMem [51], SPMem [117], and concurrent work HY-WorldPlay [37]. Yume-1.5 is a FramePack-based method that relies solely on temporal context without spatial memory. GEN3C, CaM, VMem, and SPMem condition generation on multi-view history frames to maintain memory consistency. SPMem accumulates history frames into

Figure 3: Video generation comparisons. Given a single input image from Tanks and Temples, we compare long-horizon generations (∼frame 800+) from all evaluated video models. Baselines exhibit severe quality degradation, geometric distortions, or content drifting at long horizons, while our method maintains realistic structures and appearances.

a global point cloud for conditioning. HY-WorldPlay uses discrete action control (keyboard inputs) rather than explicit camera trajectory conditioning. Since CaM and SPMem are not open-sourced, we re-implement them based on Wan2.1-14B [107].

All methods are evaluated on DL3DV-Evaluation [60] for in-domain testing and Tanks and Temples [46] for out-of-domain generalization. We follow standard protocol [81, 82, 128] and report SSIM, LPIPS, and Fréchet Inception Distance (FID). Since standard metrics are insufficient for evaluating long video generation, we additionally adopt metrics from WorldScore [18]: Subjective Quality Score for human perceptual quality, Style Consistency Score to detect visual drifting between the first and last frames, and Camera Controllability Score to measure camera pose accuracy. We further report reprojection error, computed by estimating per-frame depth with an off-the-shelf SLAM system [35], to verify 3D consistency of the generated videos.

Quantitative Comparison. As shown in Tab. 1, our method achieves the best results on both datasets across nearly all metrics, validating our anti-forgetting and anti-drifting mechanisms: 3D geometry serves as an information routing signal to enforce long-range consistency without sacrificing generation quality, while context compression and self-augmentation prevent quality degradation over long horizons. Among the baselines, each addresses only one aspect of this challenge. GEN3C [81] achieves the best Camera Controllability and Reprojection Error through explicit depth-warped conditioning, but this rigid geometric constraint degrades generation quality, as reflected by its low Subjective Quality and SSIM. CaM [128] and SPMem [117] are the strongest competitors on quality metrics thanks to their multi-view history memory, but their implicit camera conditioning leads to substantially lower Camera Controllability. SPMem's global

Table 2: Quantitative comparison on 3D scene generation. Best results are shown in bold and second best are underlined.

Method	DL3DV				Tanks-and-Temples			
	LPIPS-P↓	LPIPS-G↓	FID↓	Subj. Qual.↑	LPIPS-P↓	LPIPS-G↓	FID↓	Subj. Qual.↑
GEN3C [81] + DAv3	0.504	0.649	99.83	11.00	0.511	0.694	125.19	5.38
Yume1.5 [70] + DAv3	0.598	0.806	121.61	0.22	0.575	0.794	113.25	0.79
CaM [128] + DAv3	0.433	0.668	94.04	12.16	0.423	0.693	94.02	9.79
VMem [51] + DAv3	0.593	0.836	206.88	2.00	0.597	0.832	211.72	3.76
SPMem [117] + DAv3	0.419	0.625	93.56	13.72	0.412	0.666	94.11	9.95
Ours + DAv3	0.413	0.603	74.39	17.02	0.409	0.648	79.36	14.42
Ours Full	0.381	0.579	65.94	20.52	0.372	0.629	72.47	18.80

point cloud conditioning also introduces geometric errors over long horizons, resulting in more pronounced drifting as reflected by lower Style Consistency. VMem [51] struggles to maintain coherence over long horizons, resulting in the weakest scores across nearly all metrics. Yume-1.5 [70] and HY-WorldPlay [37] lack explicit camera trajectory conditioning, failing to follow the specified viewpoints; HY-WorldPlay further suffers from severe temporal drifting, leading to substantial quality degradation. In contrast, our framework bridges this gap, achieving both high visual fidelity and accurate camera control simultaneously.

Qualitative Comparison. In Fig. 3, we visualize single-image to long-video generation results. The shown images correspond to approximately frame 800, illustrating the challenges baselines face in maintaining realistic content over long generation horizons. VMem exhibits severe quality degradation and structural collapse; GEN3C and Yume-1.5 suffer from geometric distortions; CaM and SPMem maintain reasonable quality but show noticeable drifting. In contrast, our method maintains realistic geometric structures and appearances with respect to the input when revisiting regions.

Distilled Model. As shown in Tab. 1, our DMD-distilled model (4 steps) achieves comparable or even slightly better per-frame quality (LPIPS, FID) compared to the full model (35 steps), while camera controllability decreases moderately due to the reduced number of denoising steps.

5.3. Evaluation on 3D Scene Generation

Baselines and Metrics. In this work, we focus on large-scale 3D scene generation. To construct competitive baselines, we pair the long video generation methods from Sec. 5.2 with Depth Anything V3 [58], a state-of-the-art 3D reconstruction model that converts videos into 3DGS. We render novel views from the reconstructed 3DGS and evaluate with FID and Subjective Quality Score. We additionally report two LPIPS variants: LPIPS-G, computed between rendered novel views and ground-truth frames, which measures overall reconstruction quality; and LPIPS-P, computed between rendered novel views and the generated video frames, which quantifies the 3D consistency of the underlying video model—a more consistent video yields a more faithful 3D reconstruction and thus lower LPIPS-P. We also compare with prior generative reconstruction methods, Lyra [2] and FantasyWorld [14], which generate short videos and lift them to 3D but are inherently limited in scene scale.

Quantitative Comparison. As shown in Tab. 2, our method achieves the best results across all metrics on both datasets. Both our variants—Ours + DAv3 and Ours Full—substantially outperform all baselines in LPIPS-G, FID, and Subjective Quality, demonstrating that the 3D consistency of our generated videos translates directly into higher-quality scene reconstructions. Furthermore, Ours Full consistently outperforms Ours + DAv3 across all metrics, validating the benefit of fine-tuning the reconstruction model on our generated scenes to improve robustness to generative artifacts. Notably, our method also achieves substantially lower LPIPS-P,

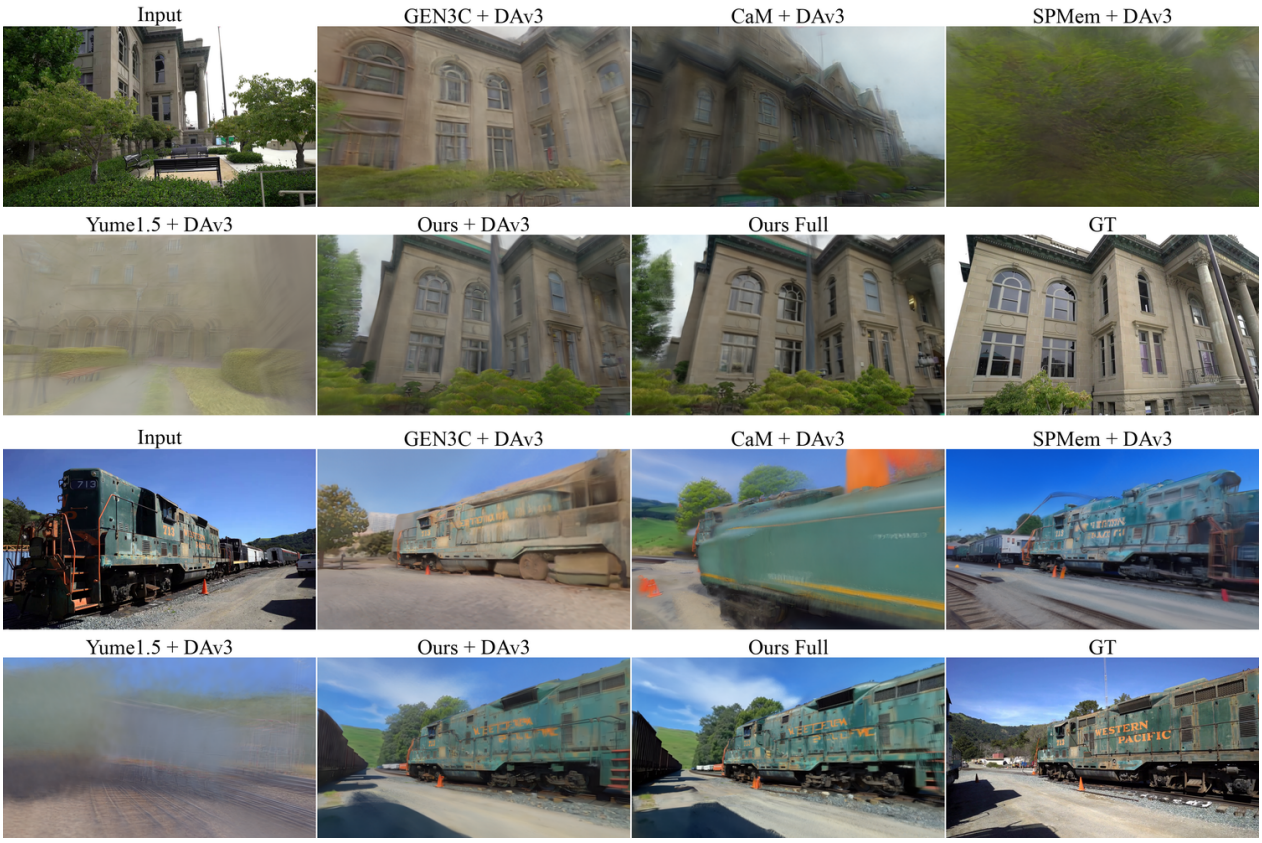

Figure 4: 3DGS comparisons. We compare renderings from 3DGS scenes reconstructed from video diffusion model outputs, starting from a single input image from Tanks and Temples.

confirming that our video model produces inherently more 3D-consistent outputs: the generated videos can be more faithfully reconstructed in 3D and re-rendered from novel viewpoints with minimal discrepancy.

Qualitative Comparison. We compare renderings of 3DGS scenes generated from single images in Fig. 4. While all baselines produce scenes with artifacts and floaters, our pipeline is able to generate realistic 3D scenes with high fidelity. We further compare with Lyra [2] and FantasyWorld [14] in Fig. 5. Both methods generate short videos and lift them to 3D, inherently limiting the achievable scene scale. In contrast, our interactive exploration framework allows users to iteratively define camera trajectories and progressively expand the environment, producing scenes of substantially greater spatial extent and complexity.

5.4. Ablation Study

We ablate the key design choices of our framework on Tanks and Temples. Quantitative results are reported in Tab. 3 and qualitative comparisons are shown in Fig. 6.

w/ Global Point Cloud fuses all history frames into a single accumulated point cloud and conditions generation on its rendered images, replacing both the per-frame 3D cache and the correspondence-based conditioning. This significantly degrades Camera Controllability (49.86 vs. 63.87) and Style Consistency (82.42 vs. 85.07), confirming that accumulated depth errors corrupt the conditioning signal over long horizons. As shown in Fig. 6, this variant produces noticeably inaccurate camera poses.

w/ Explicit Corr. Fusion replaces our learned MLP aggregation with explicit depth-reasoning-based fusion for merging correspondences from multiple source frames. Camera Controllability drops (57.29 vs. 63.87), showing that learned aggregation handles noisy depth estimates more gracefully than hard geometric fusion.

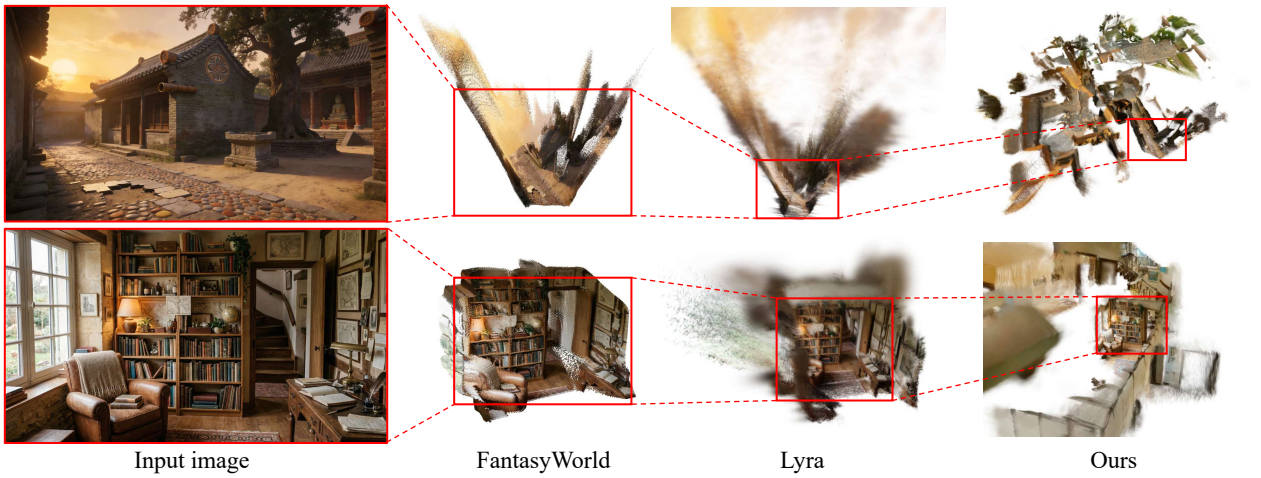

Figure 5: Qualitative comparison with Lyra and FantasyWorld. We show 3DGS renderings (Lyra and Ours) and point cloud renderings (FantasyWorld) in bird's-eye view. Red bounding boxes highlight approximately the same spatial region across methods. Our interactive exploration framework produces scenes of significantly greater scale and complexity.

Table 3: Ablation study on Tanks and Temples. We ablate key design choices of our framework. Best results are shown in bold.

Method	SSIM↑	LPIPS↓	FID↓	Subjective Qual.↑	Style Consist.↑	Camera Ctrl.↑	Reproj. Err.↓
Ours	0.384	0.552	51.33	43.35	85.07	63.87	0.069
w/ Global Point Cloud	0.368	0.562	52.54	44.58	82.42	49.86	0.067
w/ Explicit Corr. Fusion	0.370	0.554	49.13	45.71	83.28	57.29	0.071
w/o FramePack	0.362	0.549	50.98	45.27	80.61	62.62	0.079
w/o Self-Augmentation	0.363	0.568	55.15	47.88	77.98	53.92	0.066

w/o FramePack removes the FramePack temporal slots. Without temporal grounding, the model is prone to drifting, significantly reducing Style Consistency (80.61 vs. 85.07) and increasing Reprojection Error (0.079 vs. 0.069). As shown in Fig. 6, this variant exhibits pronounced visual drifting.

w/o Self-Augmentation removes the self-augmentation training strategy. While per-frame Subjective Quality improves (47.88 vs. 43.35), long-range consistency degrades substantially: Style Consistency drops to 77.98 and Camera Controllability to 53.92. Without exposure to imperfect conditioning during training, the model becomes brittle at inference when conditioning on its own imperfect outputs, causing errors to compound across segments, as visible in Fig. 6.

5.5. Applications

Beyond quantitative evaluation, we demonstrate the practical applicability of our framework through an interactive GUI, in-the-wild scene generation, and downstream simulation.

Interactive GUI. We build an interactive interface that allows users to specify camera trajectories within the 3D cache and progressively generate and explore scenes in real time, as shown in Fig. 7. The GUI visualizes the accumulated point clouds, enabling users to plan trajectories that revisit previously explored regions or venture into unobserved areas.

In-the-Wild Scene Generation. We showcase our method on diverse in-the-wild images beyond the evaluation benchmarks, generating large-scale explorable 3D scenes from a single input image. As shown in Fig. 1 and

Figure 6: Qualitative ablation study. Given a single input image, we compare generations from our full model and ablated variants on Tanks and Temples scenes.

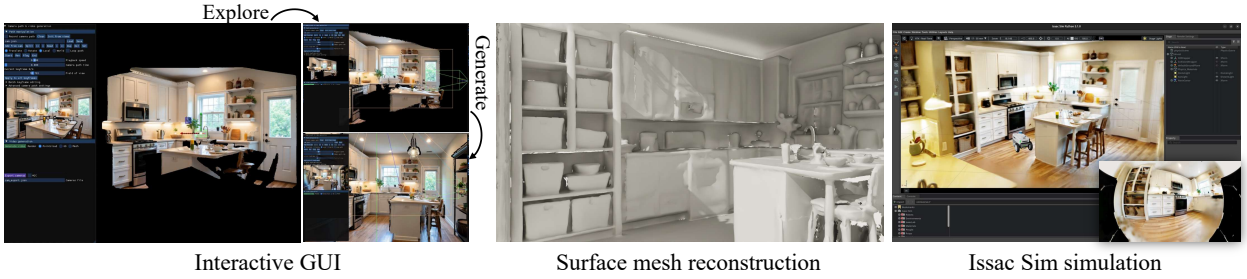

Figure 7: Applications. Our interactive interface allows users to specify camera trajectories within the 3D cache to easily generate novel viewpoints. Moreover, the reconstructed 3DGS scenes can be converted into surface meshes and integrated into embodied AI simulators such as NVIDIA Isaac Sim for robot simulation.

Fig. 8, our framework produces globally consistent long videos and high-quality 3D reconstructions across a variety of scene types, including both indoor and outdoor environments.

Embodied AI Simulation. The 3D Gaussian Splatting representations and meshes generated by our pipeline can be directly exported to physics engines for downstream applications. We demonstrate this by importing our reconstructed scenes into NVIDIA Isaac Sim, enabling physically grounded robot navigation and interaction within the generated environments. This highlights the potential of our framework for scalable embodied AI simulation without the need for real-world 3D data acquisition.

6. Discussion

In this work, we introduced Lyra 2.0, a generative reconstruction framework that enables the creation of large-scale, explorable 3D environments. Our approach addresses the key challenge of long-horizon consistency in camera-controlled video generation through dedicated anti-forgetting and anti-drifting mechanisms, and improves the reconstruction model to be robust to small generative errors. The generated scenes can be directly deployed for interactive exploration, virtual reality experiences, and simulation.

Despite these advances, several limitations remain. First, our current framework focuses on static environments and does not explicitly model dynamic scenes, which remains an important direction for future work. Second, our video generation model inherits characteristics of the training data. In particular, the DL3DV dataset contains exposure variations across views, which the model may reproduce during generation. Such photometric inconsistencies can lead to artifacts in the feed-forward 3DGS reconstruction. Addressing photometric stability within the network [16] or using photometrically consistent synthetic datasets [128], e.g., from game engines, could lead to more consistent 3D scenes.

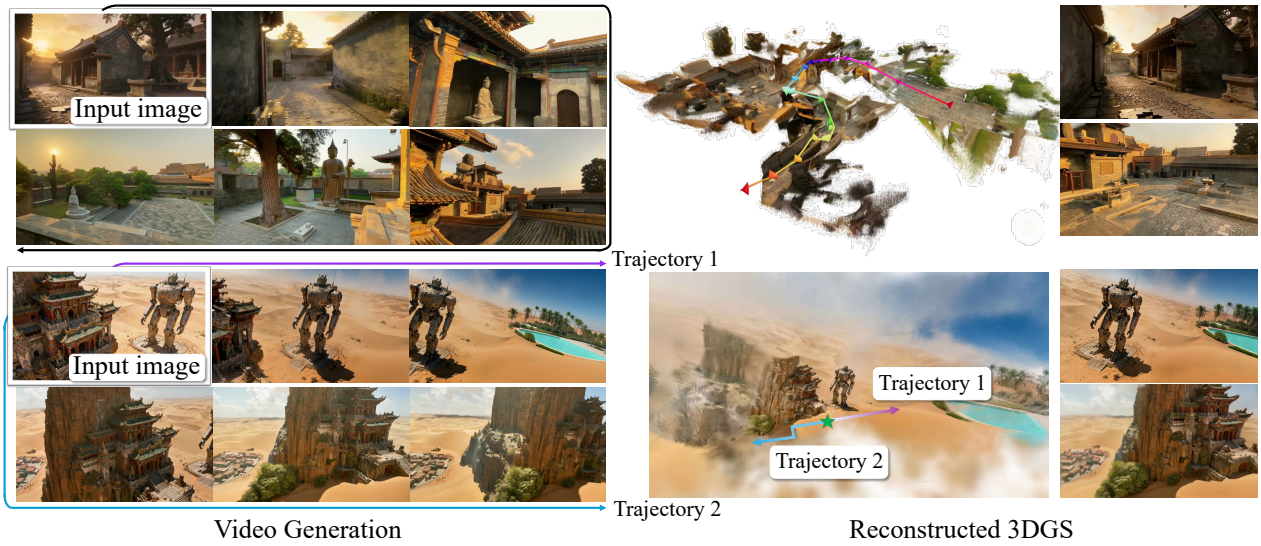

Figure 8: In-the-Wild Scene Generation. We show video generations and 3DGS reconstructions for challenging in-the-wild input images that go beyond the training data distribution. Our approach supports flexible camera trajectories specified in the GUI for world exploration, including combining multiple trajectories from the same starting point (see second example).

Acknowledgement

We would like to thank Product Managers Aditya Mahajan and Matt Cragun for their valuable guidance and support. We also thank Oliver Hahn, David Pankratz, Christian Laforte, Gene Liu, and Rafal Karp for insightful discussions and feedback. We are grateful to Yifeng Jiang, Nicolas Moenne-Loccoz, Tanki Zhang, Aditya Gupta, and Gavriel State for their prompt and helpful support in developing the Isaac Sim demo. Finally, we sincerely acknowledge Merlin Nimier-David, Thomas Müller, and Alex Keller for their foundational interactive GUI, upon which our system builds

A. Implementation Details

A.1. Model Architecture

Base Model. We build upon the Wan 2.1-14B DiT [107] as our backbone video diffusion model. The VAE encodes videos at 8× spatial and 4× temporal downsampling with a latent channel dimension C=16. All training and inference are performed at a resolution of 832×480 pixels.

Camera Conditioning Modules. We inject camera information through two complementary modules:

- Depth-warped conditioning: We forward-warp the most recent frame to each target viewpoint using the estimated depth map, encode through the VAE, and concatenate with the denoising latent along the channel dimension.
- Plücker ray injection: 6D Plücker ray coordinates are computed per pixel for all frames (temporal history, spatial memory, and generation tokens). These are projected to the DiT's hidden dimension via a pixel-shuffle layer followed by a single linear layer, yielding per-token ray embeddings $\mathbf{p}$. These are added to the token features before the query and key projections at every transformer block, i.e., $\mathbf{q} = W_Q(\mathbf{x} + \mathbf{p})$ and $\mathbf{k} = W_K(\mathbf{x} + \mathbf{p})$, while the value projection remains unmodified.

Canonical Coordinate Injection. The forward-warped 4-channel canonical coordinate maps $[\hat{\mathbf{C}}_j; \hat{D}_j]$ are downsampled to match the latent spatial resolution via pixel shuffle. Each channel is encoded with sinusoidal positional encoding, and the resulting embeddings are aggregated through a pixel-shuffle layer followed by a single linear layer. The output is injected into the queries and keys of self-attention at every transformer block, but not the values, following the same injection scheme as the Plücker ray embeddings described above. This design ensures that the correspondence signal guides which generation tokens attend to which spatial slots, while the values remain unmodified from the pretrained model.

Number of Spatial Slots. We analyze the effect of the number of retrieved spatial memory frames N_s on target-frame coverage in Fig. 9. N_s=5 provides a good trade-off between coverage of previously visited regions and inference efficiency.

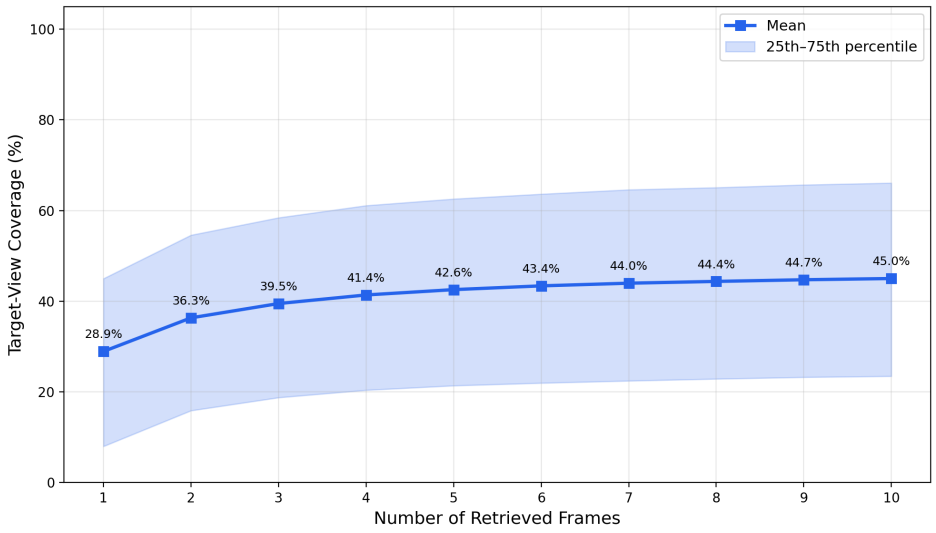

Figure 9: Target-frame coverage vs. number of retrieved spatial memory frames. We evaluate on training videos by treating the latter half as the target generation segment. Coverage is computed by forward-warping each retrieved frame to every target viewpoint using ground-truth depth; a target pixel is counted as covered only when the depth discrepancy between the warped point and the target ground-truth depth falls below a threshold. N_s=5 offers a favorable balance between spatial coverage and computational efficiency.

A.2. Training

Spatial Memory. We retrieve N_s=5 spatial memory frames per autoregressive step. The downsampled point cloud in the 3D cache uses a subsampling factor of d=8. The visibility score occlusion threshold is δ=0.1 (in normalized depth units).

Self-Augmentation. We set the augmentation probability $p_{\text{aug}} = 0.7$.

Optimization. We use AdamW [68] with a learning rate of 3×10^{-5} and weight decay 0.1. Training uses a batch size of 64 across 64 NVIDIA GB200 GPUs. We train for 7,000 iterations. All newly added modules are initialized with zero weights so that the model starts from the pretrained Wan 2.1 behavior. We use bf16 mixed-precision training throughout.

Flow Matching. We use rectified flow matching. During training we sample timesteps from a logit-normal distribution (mean 0, std 1 in logit space) with uniform time weighting; at inference we use the FlowUniPC [140] multistep scheduler with 35 steps.

A.3. Inference

Classifier-Free Guidance. We apply classifier-free guidance (CFG) with a scale of 5.0 for the text prompt.

Runtime. Each autoregressive step (80 frames) takes approximately 194 seconds on a single NVIDIA GB200 GPU for the full model (35 steps with CFG), including depth estimation, spatial memory retrieval, and DiT denoising. With Ours DMD (4 steps, no CFG), this reduces to approximately 15 seconds per step. Spatial memory retrieval takes less than 1 second per step in both cases.

A.4. 3D Reconstruction

3DGS. The Gaussian DPT head downsampling factor is k=2, reducing the Gaussian count by 4×. To construct the fine-tuning dataset, we autoregressively generate 3,000 one-minute videos using images and camera trajectories from DL3DV [60]. We then fine-tune DAv3 on these scenes for 10,000 iterations with a learning rate of 5×10^{-5} and batch size 8.

Mesh Extraction. The mesh extraction step extracts a triangular mesh of the scene using a hierarchical sparse grid. The number of levels and voxel sizes for each level in the hierarchy are determined by the scale of the scene. The depth from the Gaussian reconstruction is used to compute a signed distance field, and a single surface mesh is extracted by running marching cubes on each level and merging them at level transitions.

A.5. Related Work

We provide more extensive related work discussion in addition to Sec. 2.

3D generation. Early work on 3D generation largely focused on category-specific object synthesis, extending GAN-based frameworks to 3D by incorporating neural rendering as an inductive bias [1, 7, 17, 20, 73, 87]. The introduction of CLIP-based supervision [78] enabled more flexible generation pipelines, supporting both text-conditioned synthesis and semantic editing [11, 22, 38, 39, 83, 109]. More recently, diffusion-based methods have substantially improved visual fidelity by replacing CLIP guidance with Score Distillation Sampling (SDS) [12, 27, 42, 47, 53, 56, 57, 65, 75, 94, 110, 113, 125, 131]. To improve geometric consistency, a number of approaches explicitly enforce multi-view coherence by generating or supervising across multiple viewpoints [19, 21, 29, 41, 44, 59, 63, 64, 79, 91, 102, 106, 112, 132]. In parallel, some methods formulate scene generation as an iterative inpainting process to progressively expand 3D environments [30, 92]. Another line of work lifts 2D observations into 3D representations using NeRF [71], 3D Gaussian Splatting [43], or mesh-based formulations in combination with diffusion priors [8, 23, 66, 67, 69, 77, 96, 101, 104, 108, 127].

Feed-forward 3D models. A complementary line of research focuses on feed-forward architectures that directly infer 3D structure from images or text in a single pass [25, 34, 40, 49, 54, 76, 86, 95, 97, 98, 99, 105, 119, 121, 122, 124, 133, 136]. While these methods enable efficient 3D generation, they are generally restricted to static scene representations. Other approaches specialize in narrow domains such as facial reconstruction [45]. Some works [55, 123] address real-world dynamic scenes, but struggle to generalize to diverse generated content or large viewpoint variations.

References

[1] S. Bahmani, J. J. Park, D. Paschalidou, X. Yan, G. Wetzstein, L. Guibas, and A. Tagliasacchi. CC3D: Layout-conditioned generation of compositional 3D scenes. In Proc. ICCV, 2023. 17

[2] S. Bahmani, T. Shen, J. Ren, J. Huang, Y. Jiang, H. Turki, A. Tagliasacchi, D. B. Lindell, Z. Gojcic, S. Fidler, H. Ling, J. Gao, and X. Ren. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In ICLR, 2026. 2, 4, 8, 11, 12

[3] S. Bahmani, I. Skorokhodov, G. Qian, A. Siarohin, W. Menapace, A. Tagliasacchi, D. B. Lindell, and S. Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. Proc. CVPR, 2025. 3

[4] S. Bahmani, I. Skorokhodov, A. Siarohin, W. Menapace, G. Qian, M. Vasilkovsky, H.-Y. Lee, C. Wang, J. Zou, A. Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. Proc. ICLR, 2025. 3

[5] P. J. Ball, J. Bauer, F. Belletti, B. Brownfield, A. Ephrat, S. Fruchter, A. Gupta, K. Holsheimer, A. Holynski, J. Hron, et al. Genie 3: A new frontier for world models. Google DeepMind Blog, pages 253–279, 2025. 3

[6] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. OpenAI technical reports, 2024. 2

[7] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. De Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In Proc. CVPR, 2022. 17

[8] E. R. Chan, K. Nagano, M. A. Chan, A. W. Bergman, J. J. Park, A. Levy, M. Aittala, S. De Mello, T. Karras, and G. Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In Proc. ICCV, 2023. 17

[9] D. Charatan, S. L. Li, A. Tagliasacchi, and V. Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In Proc. CVPR, 2024. 4

[10] E. M. Chen, S. Holalkere, R. Yan, K. Zhang, and A. Davis. Ray conditioning: Trading photo-consistency for photo-realism in multi-view image generation. In ICCV, 2023. 3

[11] K. Chen, C. B. Choy, M. Savva, A. X. Chang, T. Funkhouser, and S. Savarese. Text2Shape: Generating shapes from natural language by learning joint embeddings. In Proc. ACCV, 2018. 17

[12] R. Chen, Y. Chen, N. Jiao, and K. Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. arXiv preprint arXiv:2303.13873, 2023. 17

[13] T. Cosmos. Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575, 2025. 2, 4

[14] Y. Dai, F. Jiang, C. Wang, M. Xu, and Y. Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In ICLR, 2026. 11, 12

[15] K. Dalal, D. Koceja, J. Xu, Y. Zhao, S. Han, K. C. Cheung, J. Kautz, Y. Choi, Y. Sun, and X. Wang. One-minute video generation with test-time training. In CVPR, 2025. 4

[16] I. Deutsch, N. Moënne-Loccoz, G. State, and Z. Gojcic. Ppisp: Physically-plausible compensation and control of photometric variations in radiance field reconstruction. arXiv preprint arXiv:2601.18336, 2026. 14

[17] T. DeVries, M. A. Bautista, N. Srivastava, G. W. Taylor, and J. M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In Proc. ICCV, 2021. 17

[18] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu. Worldscore: A unified evaluation benchmark for world generation. In ICCV, 2025. 10

[19] Q. Feng, Z. Xing, Z. Wu, and Y.-G. Jiang. FDGaussian: Fast Gaussian splatting from single image via geometric-aware diffusion model. arXiv preprint arXiv:2403.10242, 2024. 17

[20] J. Gao, T. Shen, Z. Wang, W. Chen, K. Yin, D. Li, O. Litany, Z. Gojcic, and S. Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. In Proc. NeurIPS, 2022. 17

[21] R. Gao, A. Holynski, P. Henzler, A. Brussee, R. Martin-Brualla, P. Srinivasan, J. T. Barron, and B. Poole. Cat3d: Create anything in 3d with multi-view diffusion models. In Proc. NeurIPS, 2024. 17

[22] W. Gao, N. Aigerman, T. Groueix, V. Kim, and R. Hanocka. TextDeformer: Geometry manipulation using text guidance. In SIGGRAPH, 2023. 17

[23] J. Gu, A. Trevithick, K.-E. Lin, J. M. Susskind, C. Theobalt, L. Liu, and R. Ramamoorthi. NerfDiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion. In Proc. ICML, 2023. 17

[24] Y. Gu, W. Mao, and M. Z. Shou. Long-context autoregressive video modeling with next-frame prediction. arXiv preprint arXiv:2503.19325, 2025. 4

[25] J. Han, F. Kokkinos, and P. Torr. VFusion3D: Learning scalable 3D generative models from video diffusion models. arXiv preprint arXiv:2403.12034, 2024. 18

[26] H. He, Y. Xu, Y. Guo, G. Wetzstein, B. Dai, H. Li, and C. Yang. Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint arXiv:2404.02101, 2024. 2, 3, 5

[27] X. He, J. Chen, S. Peng, D. Huang, Y. Li, X. Huang, C. Yuan, W. Ouyang, and T. He. GVGEN: Text-to-3D generation with volumetric representation. arXiv preprint arXiv:2403.12957, 2024. 17

[28] X. He, C. Peng, Z. Liu, B. Wang, Y. Zhang, Q. Cui, F. Kang, B. Jiang, M. An, Y. Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. arXiv preprint arXiv:2508.13009, 2025. 3

[29] L. Höllein, A. Božič, N. Müller, D. Novotny, H.-Y. Tseng, C. Richardt, M. Zollhöfer, and M. Nießner. ViewDiff: 3D-consistent image generation with text-to-image models. In Proc. CVPR, 2024. 17

[30] L. Höllein, A. Cao, A. Owens, J. Johnson, and M. Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In Proc. ICCV, 2023. 17

[31] L. Höllein and M. Nießner. World reconstruction from inconsistent views. arXiv preprint arXiv:2603.16736, 2026. 4

[32] Y. Hong, B. Liu, M. Wu, Y. Zhai, K.-W. Chang, L. Li, K. Lin, C.-C. Lin, J. Wang, Z. Yang, et al. Slowfast-vgen: Slow-fast learning for action-driven long video generation. arXiv preprint arXiv:2410.23277, 2024. 4

[33] Y. Hong, Y. Mei, C. Ge, Y. Xu, Y. Zhou, S. Bi, Y. Hold-Geoffroy, M. Roberts, M. Fisher, E. Shechtman, et al. Relic: Interactive video world model with long-horizon memory. arXiv preprint arXiv:2512.04040, 2025. 4

[34] Y. Hong, K. Zhang, J. Gu, S. Bi, Y. Zhou, D. Liu, F. Liu, K. Sunkavalli, T. Bui, and H. Tan. LRM: Large reconstruction model for single image to 3D. In Proc. ICLR, 2024. 18

[35] J. Huang, Q. Zhou, H. Rabeti, A. Korovko, H. Ling, X. Ren, T. Shen, J. Gao, D. Slepichev, C.-H. Lin, et al. Vipe: Video pose engine for 3d geometric perception. arXiv preprint arXiv:2508.10934, 2025. 9, 10

[36] X. Huang, Z. Li, G. He, M. Zhou, and E. Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. arXiv preprint arXiv:2506.08009, 2025. 8

[37] T. HunyuanWorld. Hy-world 1.5: A systematic framework for interactive world modeling with real-time latency and geometric consistency. arXiv preprint, 2025. 9, 11

[38] A. Jain, B. Mildenhall, J. T. Barron, P. Abbeel, and B. Poole. Zero-shot text-guided object generation with dream fields. In Proc. CVPR, 2022. 17

[39] N. Jetchev. ClipMatrix: Text-controlled creation of 3D textured meshes. arXiv preprint arXiv:2109.12922, 2021. 17

[40] L. Jiang and L. Wang. Brightdreamer: Generic 3D Gaussian generative framework for fast text-to-3D synthesis. arXiv preprint arXiv:2403.11273, 2024. 18

[41] Y. Kant, E. Weber, J. K. Kim, R. Khirodkar, S. Zhaoen, J. Martinez, I. Gilitschenski, S. Saito, and T. Bagautdinov. Pippo: High-resolution multi-view humans from a single image. In Proc. CVPR, 2025. 17

[42] O. Katzir, O. Patashnik, D. Cohen-Or, and D. Lischinski. Noise-free score distillation. In Proc. ICLR, 2024. 17

[43] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis. 3d gaussian splatting for real-time radiance field rendering. In ACM TOG, 2023. 4, 17

[44] S. W. Kim, B. Brown, K. Yin, K. Kreis, K. Schwarz, D. Li, R. Rombach, A. Torralba, and S. Fidler. NeuralFieldLDM: Scene generation with hierarchical latent diffusion models. In Proc. CVPR, 2023. 17

[45] T. Kirschstein, J. Romero, A. Sevastopolsky, M. Nießner, and S. Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. arXiv preprint arXiv:2502.20220, 2025. 18

[46] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. ACMTOG, 36(4), 2017. 10

[47] K. Lee, K. Sohn, and J. Shin. DreamFlow: High-quality text-to-3D generation by approximating probability flow. In Proc. ICLR, 2024. 17

[48] G. Li, S. Zheng, S. Xu, J. Chen, B. Li, X. Hu, L. Zhao, and P.-T. Jiang. Magicworld: Interactive geometry-driven video world exploration. arXiv preprint arXiv:2511.18886, 2025. 3, 4

[49] J. Li, H. Tan, K. Zhang, Z. Xu, F. Luan, Y. Xu, Y. Hong, K. Sunkavalli, G. Shakhnarovich, and S. Bi. Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model. In Proc. ICLR, 2024. 18

[50] J. Li, J. Tang, Z. Xu, L. Wu, Y. Zhou, S. Shao, T. Yu, Z. Cao, and Q. Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. arXiv preprint arXiv:2506.17201, 2025. 3

[51] R. Li, P. Torr, A. Vedaldi, and T. Jakab. Vmem: Consistent interactive video scene generation with surfel-indexed view memory. In ICCV, pages 25690–25699, 2025. 4, 9, 11

[52] X. Li, T. Wang, Z. Gu, S. Zhang, C. Guo, and L. Cao. Flashworld: High-quality 3d scene generation within seconds. In ICLR, 2026. 4

[53] Z. Li, Y. Chen, L. Zhao, and P. Liu. Controllable text-to-3D generation via surface-aligned Gaussian splatting. arXiv preprint arXiv:2403.09981, 2024. 17

[54] H. Liang, J. Cao, V. Goel, G. Qian, S. Korolev, D. Terzopoulos, K. N. Plataniotis, S. Tulyakov, and J. Ren. Wonderland: Navigating 3d scenes from a single image. Proc. CVPR, 2025. 4, 18

[55] H. Liang, J. Ren, A. Mirzaei, A. Torralba, Z. Liu, I. Gilitschenski, S. Fidler, C. Oztireli, H. Ling, Z. Gojcic, and J. Huang. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. Proc. NeurIPS, 2025. 18

[56] Y. Liang, X. Yang, J. Lin, H. Li, X. Xu, and Y. Chen. Luciddreamer: Towards high-fidelity text-to-3D generation via interval score matching. arXiv preprint arXiv:2311.11284, 2023. 17

[57] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin. Magic3D: High-resolution text-to-3D content creation. In Proc. CVPR, 2023. 17

[58] H. Lin, S. Chen, J. Liew, D. Y. Chen, Z. Li, G. Shi, J. Feng, and B. Kang. Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647, 2025. 3, 4, 6, 8, 9, 11

[59] Y. Lin, H. Han, C. Gong, Z. Xu, Y. Zhang, and X. Li. Consistent123: One image to highly consistent 3D asset using case-aware diffusion priors. In arXiv preprint arXiv:2309.17261, 2023. 17

[60] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In CVPR, pages 22160–22169, 2024. 9, 10, 17

[61] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022. 4

[62] P. Liu, Z. Guo, M. Warke, S. Chintala, C. Paxton, N. M. M. Shafiullah, and L. Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. In ICRA, 2025. 4

[63] P. Liu, Y. Wang, F. Sun, J. Li, H. Xiao, H. Xue, and X. Wang. Isotropic3D: Image-to-3D generation based on a single clip embedding. arXiv preprint arXiv:2403.10395, 2024. 17

[64] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In Proc. ICCV, 2023. 17

[65] X. Liu, X. Zhan, J. Tang, Y. Shan, G. Zeng, D. Lin, X. Liu, and Z. Liu. HumanGaussian: Text-driven 3D human generation with Gaussian splatting. In Proc. CVPR, 2024. 17

[66] Y. Liu, C. Lin, Z. Zeng, X. Long, L. Liu, T. Komura, and W. Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In Proc. ICLR, 2024. 17

[67] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In Proc. CVPR, 2024. 17

[68] I. Loshchilov. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 17

[69] Y. Lu, X. Ren, J. Yang, T. Shen, Z. Wu, J. Gao, Y. Wang, S. Chen, M. Chen, S. Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. arXiv preprint arXiv:2412.03934, 2024. 4, 17

[70] X. Mao, Z. Li, C. Li, X. Xu, K. Ying, T. He, J. Pang, Y. Qiao, and K. Zhang. Yume-1.5: A text-controlled interactive world generation model. arXiv preprint arXiv:2512.22096, 2025. 3, 9, 11

[71] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Proc. ECCV, 2020. 17

[72] K. Museth. Vdb: High-resolution sparse volumes with dynamic topology. ACM Transactions on Graphics (TOG), 32(3):1–22, 2013. 8

[73] R. Or-El, X. Luo, M. Shan, E. Shechtman, J. J. Park, and I. Kemelmacher-Shlizerman. StyleSDF: High-resolution 3D-consistent image and geometry generation. In Proc. CVPR, 2022. 17

[74] R. Po, E. R. Chan, C. Chen, and G. Wetzstein. Bagger: Backwards aggregation for mitigating drift in autoregressive video diffusion models. arXiv preprint arXiv:2512.12080, 2025. 4

[75] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In Proc. ICLR, 2023. 17

[76] G. Qian, J. Cao, A. Siarohin, Y. Kant, C. Wang, M. Vasilkovsky, H.-Y. Lee, Y. Fang, I. Skorokhodov, P. Zhuang, et al. Atom: Amortized text-to-mesh using 2d diffusion. arXiv preprint arXiv:2402.00867, 2024. 18

[77] G. Qian, J. Mai, A. Hamdi, J. Ren, A. Siarohin, B. Li, H.-Y. Lee, I. Skorokhodov, P. Wonka, S. Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In Proc. ICLR, 2024. 17

[78] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In Proc. ICML, 2021. 17

[79] X. Ren, J. Huang, X. Zeng, K. Museth, S. Fidler, and F. Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In Proc. CVPR, 2024. 17

[80] X. Ren, Y. Lu, H. Liang, Z. Wu, H. Ling, M. Chen, S. Fidler, F. Williams, and J. Huang. Scube: Instant large-scale scene reconstruction using voxsplats. Proc. NeurIPS, 2024. 4

[81] X. Ren, T. Shen, J. Huang, H. Ling, Y. Lu, M. Nimier-David, T. Müller, A. Keller, S. Fidler, and J. Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In CVPR, pages 6121–6132, 2025. 2, 3, 5, 9, 10, 11

[82] X. Ren and X. Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 10

[83] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, M. Fumero, and K. R. Malekshan. CLIP-Forge: Towards zero-shot text-to-shape generation. In Proc. CVPR, 2022. 17

[84] N. Savov, N. Kazemi, D. Zhang, D. P. Paudel, X. Wang, and L. Van Gool. Statespacediffuser: Bringing long context to diffusion world models. arXiv preprint arXiv:2505.22246, 2025. 4

[85] M.-A. Schneider, L. Höllein, and M. Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. arXiv preprint arXiv:2506.01799, 2025. 4

[86] K. Schwarz, N. Mueller, and P. Kontschieder. Generative gaussian splatting: Generating 3d scenes with video diffusion priors. arXiv preprint arXiv:2503.13272, 2025. 18

[87] K. Schwarz, A. Sauer, M. Niemeyer, Y. Liao, and A. Geiger. VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids. In Proc. NeurIPS, 2022. 17

[88] J. Seo, K. Fukuda, T. Shibuya, T. Narihira, N. Murata, S. Hu, C.-H. Lai, S. Kim, and Y. Mitsufuji. Genwarp: Single image to novel views with semantic-preserving generative warping. NeurIPS, 2024. 3

[89] A. Shabanov, P. Hedman, E. Weber, Z. Li, D. Rozumny, G. L. Lan, N. Dhingra, L. Luo, A. Vedaldi, C. Richardt, et al. Free-range gaussians: Non-grid-aligned generative 3d gaussian reconstruction. arXiv preprint arXiv:2604.04874, 2026. 4

[90] A. Sharma, A. Yu, A. Razavi, A. Toor, A. Pierson, A. Gupta, A. Waters, D. Tanis, D. Erhan, E. Lau, E. Shaw, G. Barth-Maron, G. Shaw, H. Zhang, H. Nandwani, H. Moraldo, H. Kim, I. Blok, J. Bauer, J. Donahue, J. Chung, K. Mathewson, K. David, L. Espeholt, M. van Zee, M. McGill, M. Narasimhan, M. Wang, M. Bińkowski, M. Babaeizadeh, M. T. Saffar, N. Pezzotti, P.-J. Kindermans, P. Rane, R. Hornung, R. Riachi, R. Villegas, R. Qian, S. Dieleman, S. Zhang, S. Cabi, S. Luo, S. Fruchter, S. Nørly, S. Srinivasan, T. Pfaff, T. Hume, V. Verma, W. Hua, W. Zhu, X. Yan, X. Wang, Y. Kim, Y. Du, and Y. Chen. Veo, 2024. 2

[91] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang. MVDream: Multi-view diffusion for 3D generation. In Proc. ICLR, 2024. 17

[92] J. Shriram, A. Trevithick, L. Liu, and R. Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. arXiv preprint arXiv:2404.07199, 2024. 17

[93] V. Sitzmann, S. Rezchikov, B. Freeman, J. Tenenbaum, and F. Durand. Light field networks: Neural scene representations with single-evaluation rendering. In Proc. NeurIPS, 2021. 3

[94] J. Sun, B. Zhang, R. Shao, L. Wang, W. Liu, Z. Xie, and Y. Liu. DreamCraft3D: Hierarchical 3D generation with bootstrapped diffusion prior. In Proc. ICLR, 2024. 17

[95] S. Szymanowicz, E. Insafutdinov, C. Zheng, D. Campbell, J. F. Henriques, C. Rupprecht, and A. Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. Proc. 3DV, 2025. 4, 18

[96] S. Szymanowicz, C. Rupprecht, and A. Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In Proc. ICCV, 2023. 17

[97] S. Szymanowicz, C. Rupprecht, and A. Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In Proc. CVPR, 2024. 18

[98] S. Szymanowicz, J. Y. Zhang, P. Srinivasan, R. Gao, A. Brussee, A. Holynski, R. Martin-Brualla, J. T. Barron, and P. Henzler. Bolt3d: Generating 3d scenes in seconds. arXiv preprint arXiv:2503.14445, 2025. 4, 18

[99] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. Proc. ECCV, 2024. 18

[100] J. Tang, J. Liu, J. Li, L. Wu, H. Yang, P. Zhao, S. Gong, X. Yuan, S. Shao, and Q. Lu. Hunyuan-gamecraft-2: Instruction-following interactive game world model. arXiv preprint arXiv:2511.23429, 2025. 3

[101] J. Tang, T. Wang, B. Zhang, T. Zhang, R. Yi, L. Ma, and D. Chen. Make-it-3D: High-fidelity 3D creation from a single image with diffusion prior. arXiv preprint arXiv:2303.14184, 2023. 17

[102] Z. Tang, P. Zhuang, C. Wang, A. Siarohin, Y. Kant, A. Schwing, S. Tulyakov, and H.-Y. Lee. Pixel-aligned multi-view generation with depth guided decoder. arXiv preprint arXiv:2408.14016, 2024. 17

[103] Q. Team. Qwen3-vl technical report. arXiv preprint arXiv:2511.21631, 2025. 9

[104] A. Tewari, T. Yin, G. Cazenavette, S. Rezchikov, J. Tenenbaum, F. Durand, B. Freeman, and V. Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In Proc. NeurIPS, 2023. 17

[105] D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao. Triposr: Fast 3D object reconstruction from a single image. arXiv preprint arXiv:2403.02151, 2024. 18

[106] V. Voleti, C.-H. Yao, M. Boss, A. Letts, D. Pankratz, D. Tochilkin, C. Laforte, R. Rombach, and V. Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. arXiv preprint arXiv:2403.12008, 2024. 17

[107] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, et al. Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314, 2025. 2, 4, 5, 10, 16

[108] Z. Wan, D. Paschalidou, I. Huang, H. Liu, B. Shen, X. Xiang, J. Liao, and L. Guibas. CAD: Photorealistic 3D generation via adversarial distillation. In Proc. CVPR, 2024. 17

[109] C. Wang, M. Chai, M. He, D. Chen, and J. Liao. Clip-NeRF: Text-and-image driven manipulation of neural radiance fields. In Proc. CVPR, 2022. 17

[110] H. Wang, X. Du, J. Li, R. A. Yeh, and G. Shakhnarovich. Score Jacobian chaining: Lifting pretrained 2d diffusion models for 3D generation. In Proc. CVPR, 2023. 17

[111] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud. Dust3r: Geometric 3d vision made easy. In Proc. CVPR, 2024. 4

[112] Y. Wang, M. Zhao, A. Mahdavi-Amiri, and H. Zhang. Act-r: Adaptive camera trajectories for single view 3d reconstruction. arXiv preprint arXiv:2505.08239, 2025. 17

[113] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu. ProlificDreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In Proc. NeurIPS, 2023. 17

[114] Z. Wang, Z. Yuan, X. Wang, T. Chen, M. Xia, P. Luo, and Y. Shan. Motionctrl: A unified and flexible motion controller for video generation. In SIGGRAPH, 2024. 3

[115] F. Williams, J. Huang, J. Swartz, G. Klar, V. Thakkar, M. Cong, X. Ren, R. Li, C. Fuji-Tsang, S. Fidler, et al. fvdb: A deep-learning framework for sparse, large scale, and high performance spatial intelligence. ACM Transactions on Graphics (TOG), 43(4):1–15, 2024. 8

[116] H. Wu, D. Wu, T. He, J. Guo, Y. Ye, Y. Duan, and J. Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling. arXiv preprint arXiv:2507.07982, 2025. 4

[117] T. Wu, S. Yang, R. Po, Y. Xu, Z. Liu, D. Lin, and G. Wetzstein. Video world models with long-term spatial memory. arXiv preprint arXiv:2506.05284, 2025. 3, 9, 10, 11

[118] Z. Xiao, Y. Lan, Y. Zhou, W. Ouyang, S. Yang, Y. Zeng, and X. Pan. Worldmem: Long-term consistent world simulation with memory. arXiv preprint arXiv:2504.12369, 2025. 4, 7

[119] K. Xie, J. Lorraine, T. Cao, J. Gao, J. Lucas, A. Torralba, S. Fidler, and X. Zeng. LATTE3D: Large-scale amortized text-to-enhanced3D synthesis. In Proc. ECCV, 2024. 18

[120] D. Xu, W. Nie, C. Liu, S. Liu, J. Kautz, Z. Wang, and A. Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. arXiv preprint arXiv:2406.02509, 2024. 3

[121] Y. Xu, Z. Shi, W. Yifan, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein. GRM: Large Gaussian reconstruction model for efficient 3D reconstruction and generation. In Proc. ECCV, 2024. 18

[122] Y. Xu, H. Tan, F. Luan, S. Bi, P. Wang, J. Li, Z. Shi, K. Sunkavalli, G. Wetzstein, Z. Xu, et al. DMV3D: Denoising multi-view diffusion using 3D large reconstruction model. In Proc. ICLR, 2024. 18

[123] Z. Xu, Z. Li, Z. Dong, X. Zhou, R. Newcombe, and Z. Lv. 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. arXiv preprint arXiv:2506.08015, 2025. 18

[124] Z. Yang, W. Ge, Y. Li, J. Chen, H. Li, M. An, F. Kang, H. Xue, B. Xu, Y. Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. arXiv preprint arXiv:2508.08086, 2025. 18

[125] J. Ye, F. Liu, Q. Li, Z. Wang, Y. Wang, X. Wang, Y. Duan, and J. Zhu. DreamReward: Text-to-3D generation with human preference. arXiv preprint arXiv:2403.14613, 2024. 17

[126] T. Yin, M. Gharbi, R. Zhang, E. Shechtman, F. Durand, W. T. Freeman, and T. Park. One-step diffusion with distribution matching distillation. In CVPR, 2024. 9

[127] P. Yoo, J. Guo, Y. Matsuo, and S. S. Gu. DreamSparse: Escaping from Plato's cave with 2D diffusion model given sparse views. In arXiv preprint arXiv:2306.03414, 2023. 17

[128] J. Yu, J. Bai, Y. Qin, Q. Liu, X. Wang, P. Wan, D. Zhang, and X. Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In SIGGRAPH Asia, pages 1–11, 2025. 4, 7, 9, 10, 11, 14

[129] M. YU, W. Hu, J. Xing, and Y. Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. Proc. ICCV, 2025. 3

[130] W. Yu, J. Xing, L. Yuan, W. Hu, X. Li, Z. Huang, X. Gao, T.-T. Wong, Y. Shan, and Y. Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048, 2024. 3

[131] X. Yu, Y.-C. Guo, Y. Li, D. Liang, S.-H. Zhang, and X. Qi. Text-to-3D with classifier score distillation. arXiv preprint arXiv:2310.19415, 2023. 17

[132] Y. Yuan, X. Wang, Y. Sheng, P. Chennuri, X. Zhang, and S. Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In Proc. CVPR, 2025. 17

[133] B. Zhang, T. Yang, Y. Li, L. Zhang, and X. Zhao. Compress3D: a compressed latent space for 3D generation from a single image. arXiv preprint arXiv:2403.13524, 2024. 18

[134] K. Zhang, S. Bi, H. Tan, Y. Xiangli, N. Zhao, K. Sunkavalli, and Z. Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In Proc. ECCV, 2024. 4

[135] L. Zhang, S. Cai, M. Li, G. Wetzstein, and M. Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In NeurIPS, 2025. 3, 4, 5, 7

[136] S. Zhang, H. Xu, S. Guo, Z. Xie, H. Bao, W. Xu, and C. Zou. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. arXiv preprint arXiv:2505.11992, 2025. 18

[137] T. Zhang, S. Bi, Y. Hong, K. Zhang, F. Luan, S. Yang, K. Sunkavalli, W. T. Freeman, and H. Tan. Test-time training done right. arXiv preprint arXiv:2505.23884, 2025. 4

[138] Y. Zhang, C. Peng, B. Wang, P. Wang, Q. Zhu, F. Kang, B. Jiang, Z. Gao, E. Li, Y. Liu, et al. Matrix-game: Interactive world foundation model. arXiv preprint arXiv:2506.18701, 2025. 3

[139] J. Zhao, F. Wei, Z. Liu, H. Zhang, C. Xu, and Y. Lu. Spatia: Video generation with updatable spatial memory. arXiv preprint arXiv:2512.15716, 2025. 2, 3, 4

[140] W. Zhao, L. Bai, Y. Rao, J. Zhou, and J. Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In NeurIPS, 2023. 17

[141] S. Zhou, Y. Du, Y. Yang, L. Han, P. Chen, D.-Y. Yeung, and C. Gan. Learning 3d persistent embodied world models. arXiv preprint arXiv:2505.05495, 2025. 2, 4