LYRA: GENERATIVE 3D SCENE RECONSTRUCTION VIA VIDEO DIFFUSION MODEL SELF-DISTILLATION

Sherwin Bahmani[1,2,3] **Tianchang Shen**[1,2,3] **Jiawei Ren**[1] **Jiahui Huang**[1] **Yifeng Jiang**[1] **Haithem Turki**[1] **Andrea Tagliasacchi**[2,4] **David B. Lindell**[2,3] **Zan Gojcic**[1] **Sanja Fidler**[1,2,3] **Huan Ling**[1] **Jun Gao**[1*] **Xuanchi Ren**[1,2,3*]

[1]NVIDIA [2]University of Toronto [3]Vector Institute [4]Simon Fraser University

*equal contribution

https://research.nvidia.com/labs/toronto-ai/lyra

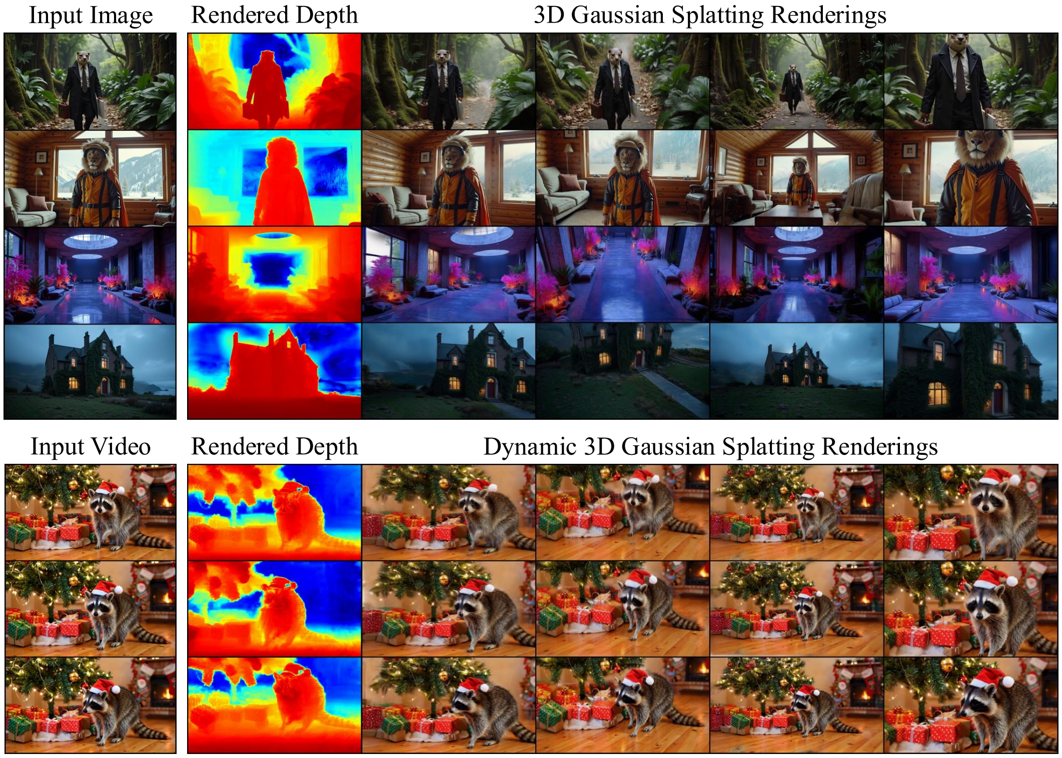

Input Image
Rendered Depth
3D Gaussian Splatting Renderings
Input Video
Rendered Depth
Dynamic 3D Gaussian Splatting Renderings

Figure 1: **Feed-Forward 3D and 4D Scene Generation.** From a single image (top), *Lyra* infers a 3D Gaussian Splatting (3DGS) representation in a feed-forward fashion, through self-distilling a video diffusion model without requiring real-world multi-view data. With a video input (bottom), *Lyra* infers a dynamic 3DGS that offers interactive control in both time (rows) and viewpoint (columns).

ABSTRACT

The ability to generate virtual environments is crucial for applications ranging from gaming to physical AI domains such as robotics, autonomous driving, and industrial AI. Current learning-based 3D reconstruction methods rely on the availability of captured real-world multi-view data, which is not always readily available. Recent advancements in video diffusion models have shown remarkable imagination capabilities, yet their 2D nature prevents their use in simulations where a robot needs to navigate and interact with the environment. In this paper, we propose a self-distillation framework that aims to distill the implicit 3D knowledge in the video diffusion models into an explicit 3D Gaussian Splatting (3DGS) representation, eliminating the need for multi-view training data. Specifically, we augment the typical RGB decoder with a 3DGS decoder, which is supervised by the output of the RGB decoder. In this approach, the 3DGS decoder can be purely trained with synthetic data generated by video diffusion models. At inference time, our model can synthesize 3D scenes from either a text prompt or a single image for real-time rendering. Our framework further extends to dynamic 3D scene generation from a monocular input video. Experimental results show that our framework achieves state-of-the-art performance in static and dynamic 3D scene generation. Video results: `https://research.nvidia.com/labs/toronto-ai/lyra`

1 Introduction

Creating high-quality 3D environments at scale is a long-standing challenge in computer vision and graphics, empowering applications across film-making, VR/AR, and physical AI with closed-loop simulation. These applications require explicit 3D representations that support real-time rendering, physical interaction, and consistent multi-view synthesis.

Recent advances in neural 3D reconstruction (Mildenhall et al., 2020; Kerbl et al., 2023; Moenne-Loccoz et al., 2024) enable recovering such representations from posed images. However, the reliance on accurate camera poses and high-quality images significantly limits their scalability. The challenge is even greater for dynamic scenes, which typically require synchronized multi-camera setups (Li et al., 2022). Reconstruction methods are also inherently constrained to the observed content and cannot extrapolate beyond the input views. Recent works aim to mitigate these limitations by using feed-forward reconstruction models (Zhang et al., 2024e; Ren et al., 2024c), but these approaches face their own bottleneck: the scarcity of diverse, large-scale 3D training data, which leads to poor out-of-domain generalization.

Video diffusion models (Cosmos, 2025; Wan, 2025) offer a promising alternative. Trained on massive internet-scale video corpora, they achieve impressive fidelity and generalization across diverse environments. By learning from real-world videos with varied camera trajectories, ranging from handheld motion to cinematic panning and drone footage, these models implicitly encode cues about the underlying 3D world without requiring multi-view training data. Unlike reconstruction approaches (Charatan et al., 2024), generative models can also hallucinate plausible content beyond what is visible in the input frames. Yet, video diffusion models generate only 2D frames, lacking explicit 3D representations. This limits their use in simulation and downstream tasks that demand geometric consistency, long-term coherence, and physical interaction. Encouragingly, recent work (Wang et al., 2024e; Ren et al., 2025) has shown that video models can be adapted for explicit camera control using relatively small-scale posed datasets. This ability to generate posed image sequences transforms them into a powerful tool that can serve as both input and supervision for 3D reconstruction models.

Motivated by this insight, we bridge these two paradigms, reconstruction and generation, through what we call *generative 3D scene reconstruction*. We introduce *Lyra*, a novel method for generating explicit 3D environments from the latent representations of video diffusion models in a single forward pass. Cruically, *Lyra* is trained in a self-distillation framework (Fig. 2), with a 3D Gaussian Splatting (3DGS) decoder operating directly in the latent space of a video diffusion model. Given a single input image or video, we sample a camera trajectory as conditioning input, denoise the resulting video latent, and decode it along two parallel branches. The first branch uses the standard RGB decoder to synthesize a video sequence, while the second is our 3DGS decoder, which produces an explicit 3D representation. Together, these branches form a self-distillation framework in which the RGB branch (teacher) supervises the 3DGS branch (student). To extend viewpoint coverage, we sample multiple camera trajectories, generate multiple video latents, and train the 3DGS decoder to fuse information across them while mitigating long-term and multi-view inconsistencies.

This generative reconstruction framework provides three key benefits: **(i)** it enables generation of large-scale synthetic environments spanning diverse scenarios directly from video diffusion models, removing the need for real-world multi-view captures; **(ii)** by operating in latent space of a video model, it allows efficient processing of multiple views without the heavy memory overhead of pixel-space feed-forward reconstruction methods; and **(iii)** its explicit 3DGS output guarantees geometric consistency, providing representations that are directly applicable to downstream tasks such as physical simulation. Consequently, at inference time, *Lyra* generates high-quality 3D Gaussian scenes from monocular input that support real-time rendering, *without requiring any additional optimization or post-processing*.

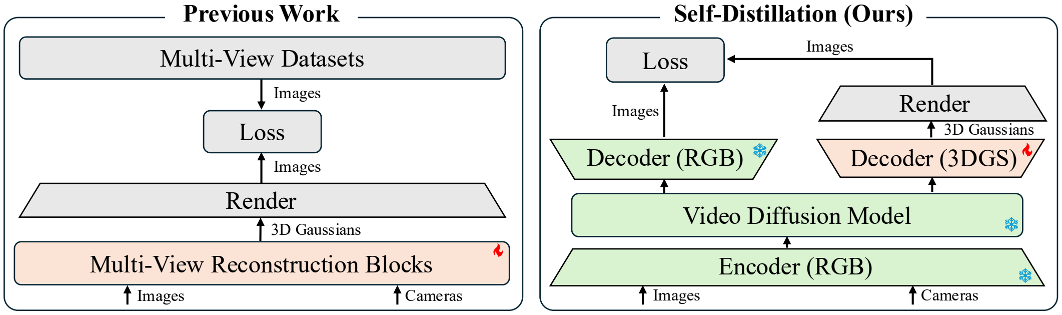


Figure 2: **Self-distillation framework of *Lyra*.** Previous work (left) (Szymanowicz et al., 2025b) trains multi-view reconstruction blocks using real-world datasets with limited diversity (Zhou et al., 2018; Ling et al., 2024b). In contrast, we propose a self-distillation framework (right) for generative 3D scene reconstruction. Precisely, a pre-trained camera-controlled video diffusion model with its RGB decoder output (teacher) supervises the rendering of the 3DGS decoder (student). Using a video model pre-trained on diverse 2D video data allows us to provide diverse multi-view supervision.

We further extend this self-distillation framework to dynamic 4D generation from monocular video. In this setting, the video model (teacher) provides space–time supervision, while the student learns to produce time-conditioned 3DGS representations that enable novel-view synthesis of dynamic scenes.

Overall, our work makes the following contributions:

- We introduce a self-distillation framework that trains a 3DGS student decoder using a pre-trained camera-controlled video diffusion model as the teacher, eliminating the need for captured multi-view real-world data.
- Our framework extends to dynamic scenes to support 4D reconstruction from a monocular video.
- Our model generalizes across diverse scenes, achieving state-of-the-art results in single-image 3D scene generation and single-video 4D scene generation.

2 Related Work

Multi-view image generation. Early works on multi-view image generation (Watson et al., 2023; Liu et al., 2023; Shi et al., 2024; Wang & Shi, 2023) mainly focus on object-centric scenes without background. Follow-up works (Sargent et al., 2024; Tang et al., 2023b; Schneider et al., 2025) extend multi-view image generators to the scene scale. CAT3D (Gao et al., 2024b) extends a pre-trained image diffusion model for multi-view image generation from any generated or real image input. In a subsequent stage, the generated multi-view images are reconstructed into an explicit 3D representation, e.g., NeRF (Mildenhall et al., 2020) or 3DGS (Kerbl et al., 2023). Moreover, these methods have been extended to generate dynamic 3D scenes from multiple viewpoints using a space-time diffusion model (Watson et al., 2024; Wu et al., 2025b; Kuang et al., 2024; Wang et al., 2025a). Even though these methods demonstrate impressive results, grounding the generations in an explicit 3D representation requires an expensive optimization stage that is not amortized across scenes. Instead of creating multi-view images and then reconstructing using optimization, we are interested in *directly* generating a 3D scene from text or a single image.

Camera-conditioned video generation. There has been significant progress in fine-tuning video diffusion models for 3D camera control. MotionCtrl (Wang et al., 2024e) pioneered camera control by conditioning pre-trained video models with camera poses. Follow-up works (He et al., 2024a; Xu et al., 2024c; Bahmani et al., 2025a;b) represent cameras as Plücker coordinates for pixel-wise conditioning. Another line of work (Hu et al., 2024; Hou et al., 2024; Zhang et al., 2024b) investigates training-free camera control of video diffusion models. Recently, SynCamMaster (Bai et al., 2025b) and ReCamMaster (Bai et al., 2025a) demonstrated impressive camera-controlled video synthesis using large-scale synthetic training data. While these works achieve compelling results, their output is not grounded in 3D. Our work is orthogonal to this line of work, as we tackle the task of feed-forward 3D generation from camera-controlled video models, allowing novel viewpoint rendering *after* generation. Concretely, we build upon GEN3C (Ren et al., 2025), a recent camera-controlled video diffusion model, and use it as a teacher within a self-distillation framework for 3D reconstruction.

Feed-forward 3D models. Early work on feed-forward 3D reconstruction mainly focused on object-centric scenes without background. These methods directly predict a NeRF (Hong et al., 2024; Li et al., 2024b) or 3DGS (Tang et al., 2024a; Szymanowicz et al., 2024) from either text or an image. More recent works (Charatan et al., 2024; Zhang et al., 2024e; Szymanowicz et al., 2025a; Ren et al., 2024c) focus on scene-level 3D reconstruction, typically regressing per-pixel 3D Gaussians from one or more input images. However, most scene-level 3D reconstruction methods are limited to training distribution scenes, e.g., RealEstate10K (Zhou et al., 2018), with limited generalizability to generated scenes. Recent methods extend feed-forward 3D reconstruction models to generate scenes. Bolt3D (Szymanowicz et al., 2025b) trains a pointmap (Wang et al., 2024b) autoencoder to generate pointmaps along the multi-view images, used for feed-forward 3D reconstruction. Closely related to our work, Wonderland (Liang et al., 2025a) uses a camera-controlled video model to generate 3D Gaussians with a feed-forward network. In contrast to these works, we use a camera-controlled video model as a teacher within a self-distillation framework to train our student 3D model *without requiring real-world multi-view data*. Importantly, we show a simple extension of our framework for feed-forward 4D scene generation, an unexplored task to date.

3 Self-Distillation using Video Diffusion Models

Our key idea is to distill the implicit 3D knowledge embedded in video diffusion models into an explicit 3DGS decoder capable of producing high-quality 3D representations. To this end, we build a teacher–student framework in which the video diffusion model (teacher) generates RGB videos that supervise the 3DGS decoder (student) that operates in the same latent space, as illustrated in Fig. 2. In this section, we detail our video diffusion model backbone (Sec. 3.1) and its role in self-distillation for the 3DGS decoder (Sec. 3.2). We discuss the design choice for the 3DGS decoder in Sec. 4 and outline minimal changes required to adapt the pipeline to dynamic 3D scenes in Sec. 5.

3.1 Background: Camera-Controlled and 3D-Consistent Video Diffusion

Since we supervise our 3DGS decoder only with the camera-controlled video diffusion model, we heavily rely on its 3D consistency. Due to this, we build our approach on GEN3C (Ren et al., 2025), a recent camera-conditioned video diffusion model. In the following we provide an overview of the key design choices made in GEN3C.

Spatiotemporal 3D cache. To improve video consistency and camera control precision, GEN3C constructs a spatiotemporal 3D cache $\{\mathbf{P}^{t,v}\}$ from input image(s) or videos, where each $\mathbf{P}^{t,v}$ is a colored point cloud obtained by unprojecting the depth estimation (Wang et al., 2024a) of an RGB image captured from camera viewpoint v at time t. The cache is organized as an $L \times V$ array, where L denotes the number of frames (temporal length) and V denotes the number of camera views.

Rendering and structured guidance. To leverage this cache, GEN3C renders each point cloud from arbitrary given camera poses, producing RGB images $\mathbf{I}^{t,v}$ and disocclusion masks $\mathbf{M}^{t,v}$ via $(\mathbf{I}^{t,v}, \mathbf{M}^{t,v}) = \mathcal{R}(\mathbf{P}^{t,v}, \mathbf{C}^t)$, where $\mathcal{R}$ denotes the rendering function that projects the 3D point cloud $\mathbf{P}^{t,v}$ onto the 2D camera plane according to the camera pose $\mathbf{C}^t$ at time t. Given a sequence of camera poses $\mathbf{C} = (\mathbf{C}^1, \ldots, \mathbf{C}^L)$, rendering all cache elements produces V videos of length L, which can be stacked into image sequences $\mathbf{I}^v \in \mathbb{R}^{L\times 3\times H\times W}$ and mask sequences $\mathbf{M}^v \in \mathbb{R}^{L\times 1\times H\times W}$ for each view v. The disocclusion masks indicate areas that the video diffusion model should fill in. These renderings serve as structured visual guidance for subsequent video generation.

Video variational autoencoder (VAE). In diffusion-based video generation models, a video variational autoencoder (VAE) (Kingma & Welling, 2013; Rombach et al., 2022) is commonly employed to compress videos into a lower-dimensional latent space for efficient training and inference. Given a RGB video $\mathbf{I} \in \mathbb{R}^{L\times 3\times H\times W}$, a pre-trained VAE encoder $\mathcal{E}$ will encode the video into a latent space, i.e. $\mathbf{z} = \mathcal{E}(\mathbf{I}) \in \mathbb{R}^{L'\times C\times h\times w}$. The training and inference of the diffusion model are then performed in this latent space. The final video $\hat{\mathbf{I}} = \mathcal{D}_{rgb}(\mathbf{z})$ is decoded with a pre-trained VAE decoder $\mathcal{D}_{rgb}$. In this paper, we adopt the pre-trained GEN3C (Ren et al., 2025; Cosmos, 2025) model. Specifically, the latent channel dimension is $C = 16$, the temporal dimension is $L' = (L - 1)/\tau + 1$, and the spatial dimensions are $h = H/\sigma$ and $w = W/\sigma$, where the temporal compression factor is $\tau = 8$ and the spatial compression factor is $\sigma = 8$.

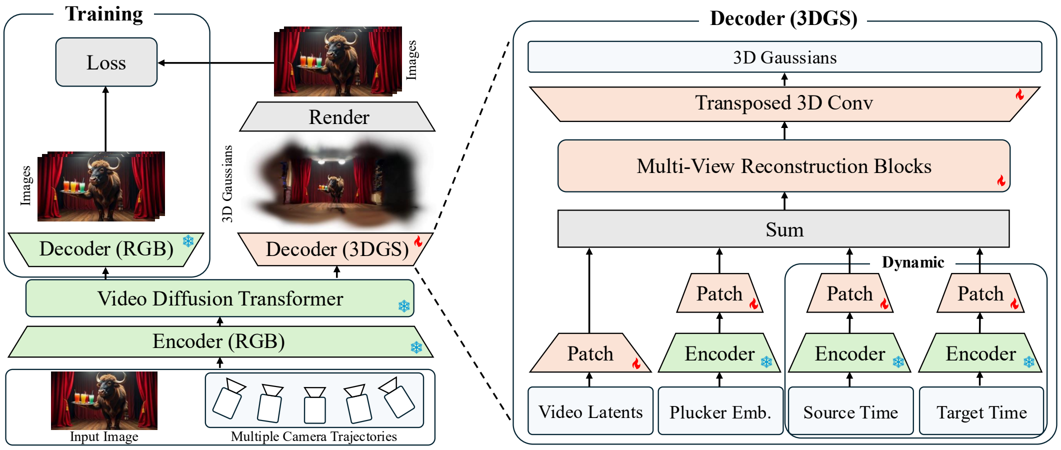


Figure 4: **3D Generative Reconstruction Framework.** Our pipeline builds upon a camera-controlled video diffusion model (Ren et al., 2025) pre-trained on large scale data. We train a 3D Gaussian Splatting (3DGS) decoder by aligning the 2D image renderings of generated 3DGS scenes with the RGB-decoded generations of the pre-trained video model. We only train the 3DGS decoder while freezing the pre-trained autoencoder and diffusion model. At inference time we directly use the 3DGS decoder, without requiring the RGB decoder anymore. Time conditioning within the 3DGS decoder allows us to easily extend our approach from static to dynamic 3D scene generation.

3.2 Self-Distillation

We train the 3DGS decoder $\mathcal{D}_s$ as a student under a teacher–student paradigm, where a camera-controlled video diffusion model $\mathcal{V}$ serves as the teacher. For diverse supervision, we curate a large-scale set of diverse text prompts with large language models (OpenAI, 2025; Bai et al., 2023), generate images I with an image diffusion model (Black Forest Labs, 2024), and expand into multi-view sequences using GEN3C (Ren et al., 2025).

Teacher–student setup. Given an input image I and a sampled camera trajectory $\{\mathbf{C}^t\}_{t=1}^{L}$, the video diffusion model $\mathcal{V}$ generates a denoised video latent $\mathbf{z} = \mathcal{V}(I, \{\mathbf{C}^t\}_{t=1}^{L})$. The latent $\mathbf{z}$ is decoded along two branches: the pre-trained RGB decoder $\mathcal{D}_{rgb}$ produces video frames $\mathbf{I}_{\mathcal{D}_{rgb}} = \mathcal{D}_{rgb}(\mathbf{z})$, while the 3DGS decoder $\mathcal{D}_s$ outputs explicit 3D Gaussians $\mathbf{G}$. Rendered views from 3DGS *(student)* are defined as $\mathbf{I}_{\mathcal{D}_s} = \mathrm{Render}(\mathbf{G}, \{\mathbf{C}^t\}_{t=1}^{L})$ and are supervised to match RGB frames $\mathbf{I}_{\mathcal{D}_{rgb}}$ *(teacher)*, forming the self-distillation loop.

Multi-trajectory supervision. In our paper, to enlarge viewpoint coverage for the input image, we sample $V = 6$ camera trajectories per input image, as shown in Fig. 3. For each trajectory v, we construct a spatiotemporal cache $\{\mathbf{P}^{t,v}\}_{t=1}^{L}$ along with $L = 121$ camera poses $\{\mathbf{C}^{t,v}\}_{t=1}^{L}$. Passing these caches through video diffusion model $\mathcal{V}$ yields latents $\mathbf{z}^v$, which the teacher decodes into RGB frames $\mathbf{I}^v_{\mathcal{D}_{rgb}} = \mathcal{D}_{rgb}(\mathbf{z}^v)$. The 3DGS decoder $\mathcal{D}_s$ learns to fuse these multiple $\mathbf{z}^v$ into coherent Gaussians $\mathbf{G}$, while filling disoccluded regions. Through this self-distillation design, the 3DGS decoder $\mathcal{D}_s$ is trained entirely from synthetic supervision provided by $\mathcal{V}$. Operating in latent space enables efficient aggregation of multiple trajectories, while the explicit Gaussian output $\mathbf{G}$ enforces geometric consistency and supports downstream simulation and real-time rendering.

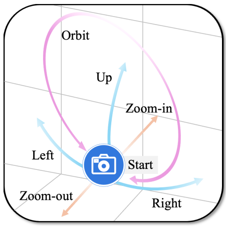


Figure 3: Sampled six camera trajectories to maximize view coverage.

4 Feed-Forward Reconstruction from Multi-View Video Latents

Our feed-forward 3DGS decoder $\mathcal{D}_s$ is designed to transform the synthesized multi-view latents generated by our video model $\mathcal{V}$ into an explicit 3D representation that can be rendered from arbitrary viewpoints. We choose 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) due to its explicit representation and fast rendering speed. We visualize our detailed 3DGS decoder in Fig. 4 (right).

4.1 3DGS Decoder

Scaling Latent-Based 3D Reconstruction. Previous feed-forward reconstruction frameworks generate 3D Gaussians for each pixel, and thus, are limited by the number and resolution of input views they can handle. For example, GS-LRM (Zhang et al., 2024e) operates on 2–4 images at 512×512 resolution, while AnySplat (Jiang et al., 2025a) is trained with 24 images at 448×448 resolution. Our video model synthesizes $V \times L = 6 \times 121 = 726$ input views at 704×1280 resolution—far beyond the capacity of existing methods, which exceeds GPU memory limit. The bottleneck lies in the attention mechanism applied to visual tokens, whose memory and compute requirements grow with the number of pixels. To overcome this limitation, we avoid scaling in pixel space and instead operate directly in the compressed video latent space produced by our camera-controlled video diffusion model. The multi-view video latents are denoted as $\mathbf{Z} \in \mathbb{R}^{V \times L' \times C \times h \times w}$.

Architecture. The 3DGS decoder $\mathcal{D}_s$ first maps its inputs into the hidden dimension of the main reconstruction blocks and outputs per-pixel 3D Gaussian features $\mathbf{G} \in \mathbb{R}^{V \times L \times H \times W \times 14}$. Specifically, the inputs are the multi-view video latents $\mathbf{Z}$ and encoded Plücker embeddings $\mathbf{E}$, both with latent dimensions $\mathbf{Z}, \mathbf{E} \in \mathbb{R}^{V \times L' \times C \times h \times w}$. Each input component is first patchified to match the hidden dimension of the main reconstruction blocks. We do not require any additional visual encoder—only a spatial 2×2 patchification layer (Dosovitskiy et al., 2021) to transform the video latents into flattened tokens for the reconstruction network. The sum of both inputs is processed through the reconstruction blocks. We follow the design introduced in Long-LRM (Ziwen et al., 2024), i.e., one block consists of one Transformer (Vaswani et al., 2017) layer followed by seven Mamba-2 (Dao & Gu, 2024) layers. We repeat the block twice to get 16 layers with 512 hidden dimensions. Finally, a transposed 3D convolution maps the hidden representation to 14 Gaussian channels: 3D position (x, y, z), scale (s_x, s_y, s_z), rotation quaternion (q_w, q_x, q_y, q_z), opacity α, and RGB (r, g, b). Formally, the static decoder is expressed as $\mathbf{G} = \mathcal{D}_s(\mathbf{Z}, \mathbf{E})$.

Plücker embeddings. Raw Plücker embeddings are first computed from camera poses $\{\mathbf{C}^{t,v}\}$ and intrinsics as $\mathbf{E}^{\text{raw}} \in \mathbb{R}^{V \times L \times 6 \times H \times W}$. We reuse the RGB encoder $\mathcal{E}$ from the pre-trained VAE to encode Plücker embeddings. Specifically, we separately encode the 3-dimensional ray directions and the 3-dimensional cross product of ray directions and origins using $\mathcal{E}$, and concatenate latent along the channel dimension to obtain $\mathbf{E}^{\text{enc}} \in \mathbb{R}^{V \times L' \times 2C \times h \times w}$. A lightweight MLP maps $\mathbf{E}^{\text{enc}}$ to $\mathbf{E} \in \mathbb{R}^{V \times L' \times C \times h \times w}$, reducing the channel dimension by a factor of 2 to match video latent $\mathbf{Z}$.

4.2 Loss function

Image-based supervision. We supervise our 3DGS decoder with an image-based reconstruction loss. The reconstruction loss is split into a Mean Squared Error (MSE) loss $\mathcal{L}_{mse}$ and an LPIPS loss $\mathcal{L}_{lpips}$ (Zhang et al., 2018) using VGG (Simonyan & Zisserman, 2014) as a feature extractor.

Depth supervision. We observed that the decoder trained with only RGB loss often produces flattened geometry. To address this, we additionally supervise the rendered depth maps using the consistent video depth estimated by an off-the-shelf system ViPE (Huang et al., 2025), and the scale-invariant depth loss $\mathcal{L}_{depth}$ from Long-LRM (Ziwen et al., 2024).

Opacity-based pruning. Similar to Long-LRM (Ziwen et al., 2024), we use an L1 regularization on the opacity $\mathcal{L}_{opacity}$ and remove the Gaussians with the lowest 80% opacity.

Total loss. Our total loss is computed as the weighted sum of all losses with weight factors λ_i

$$\mathcal{L} = \lambda_{mse}\,\mathcal{L}_{mse} + \lambda_{lpips}\,\mathcal{L}_{lpips} + \lambda_{depth}\,\mathcal{L}_{depth} + \lambda_{opacity}\,\mathcal{L}_{opacity} \tag{1}$$

where we set $\lambda_{mse} = 1.0$, $\lambda_{lpips} = 0.5$, $\lambda_{depth} = 0.05$, and $\lambda_{opacity} = 0.1$.

5 Extension to Dynamic 3D Scenes

Our approach can be extended to handle dynamic 3D scenes with minimal changes, outlined below.

Self-distillation for dynamic scenes. Our dynamic 3D setup closely follows the design of the static 3D counterpart. Instead of a single image, the video model takes a single video of length L with corresponding camera poses as input and generates multi-view video latents capturing the same underlying motion. For video inputs, we follow a protocol similar to the static case: sampling

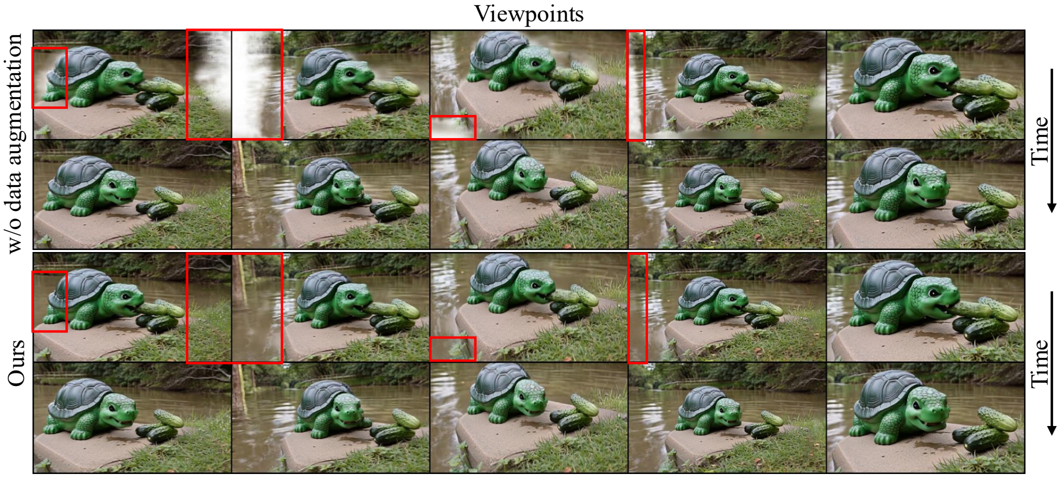


Figure 5: **Dynamic data augmentation.** When naively supervising the time-aware 3DGS decoder, we observe artifacts in the generated 3D Gaussians. Specifically, early timesteps from extreme poses exhibit low opacity in regions not covered by the supervision signal. To address this, we augment the supervision data with a motion-reversed video, ensuring that each timestep is observed from the full spatial coverage, thereby preventing low opacity artifacts in the early timesteps.

diverse text prompts with large language models, then generating videos using video diffusion models (Cosmos, 2025; Wan, 2025). The generated videos are then annotated with camera poses and depth maps using ViPE (Huang et al., 2025).

Dynamic 3DGS decoder architecture. We follow the bullet-time design of Liang et al. (2025b) and design our dynamic 3DGS decoder $\mathcal{D}_s$ to generate time-dependent 3D Gaussians for all the video frames. Specifically, the dynamic 3DGS decoder $\mathcal{D}_s$ closely follows the architecture of $\mathcal{D}_s$, except that we augment the $\mathcal{D}_s$ input with the encoded source and target time embeddings $\mathbf{T}^{\text{src}}, \mathbf{T}^{\text{tgt}} \in \mathbb{R}^{V \times L' \times C \times h \times w}$. The unencoded (raw) source times are assigned to each input frame, forming $\mathbf{T}^{\text{src,raw}} \in \mathbb{R}^{V \times L \times 1 \times H \times W}$, and the target time is $\mathbf{T}^{\text{tgt,raw}} \in \mathbb{R}^{V \times 1 \times 1 \times H \times W}$. Both are spatially repeated and augmented with a 2-dimensional sinusoidal embedding to expand the channel dimension to 3, then encoded with the RGB encoder $\mathcal{E}$ to latent dimensions $\mathbf{T}^{\text{src}}, \mathbf{T}^{\text{tgt}} \in \mathbb{R}^{V \times L' \times C \times h \times w}$. The target time is repeated along the latent temporal dimension to match L'. During training, a target time step is randomly sampled, and the corresponding 3D Gaussians are decoded and supervised across all trajectories generated with dynamic data augmentation. We fine-tune $\mathcal{D}_d$ from a pre-trained $\mathcal{D}_s$ with the patchification layers of $\mathbf{T}^{\text{src}}$ and $\mathbf{T}^{\text{tgt}}$ initialized to zeros. $\mathbf{T}^{\text{src}}$ and $\mathbf{T}^{\text{tgt}}$ are directly added to the sum of $\mathbf{Z}$ and $\mathbf{E}$. Formally, the dynamic decoder is defined as $\mathbf{G} = \mathcal{D}_d(\mathbf{Z}, \mathbf{E}, \mathbf{T}^{\text{src}}, \mathbf{T}^{\text{tgt}})$.

Dynamic data augmentation. In the static 3D scenes, where time is not a factor, frames from different timesteps can all be used to supervise the 3DGS renderings. In contrast, for dynamic 3D scenes, the 3DGS changes over time, so only frames from the corresponding timestamp are valid for supervision. This restriction can lead to a trivial solution, where the 3DGS at a given timestep simply ignores information from frames at other timesteps (see Fig. 5).

To mitigate this issue, we introduce a dynamic data augmentation strategy that balances supervision across timesteps. Early frames are naturally associated with viewpoints close to the input image, whereas later frames correspond to more distant viewpoints. To counter this imbalance, we augment the training data with paired supervision views for each timestep — one near and one far. Concretely, we reverse the input video in frame order and feed it into the video model $\mathcal{V}$, which produces six additional multi-view sequences. After reversing these sequences back to the original motion order, we obtain six trajectories that progress inward (from far viewpoints toward the input). Combined with the original six outward trajectories, this yields 12 supervision views per timestep. Importantly, this augmentation is applied only during training: the reversed trajectories act purely as extra supervision signals to prevent collapse from pruning artifacts and are not required at inference time.

Table 1: **State-of-the-art comparisons.** We compare our method with previous works for single image to 3D generation using RealEstate10K, DL3DV, and Tanks-and-Temples datasets.

Method	RealEstate10K			DL3DV			Tanks-and-Temples		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ZeroNVS	13.01	0.378	0.448	13.35	0.339	0.465	12.94	0.325	0.470
ViewCrafter	16.84	0.514	0.341	15.53	0.525	0.352	14.93	0.483	0.384
Wonderland	17.15	0.550	0.292	16.64	0.574	0.325	15.90	0.510	0.344
Bolt3D	21.54	0.747	0.234	-	-	-	-	-	-
Ours	**21.79**	**0.752**	**0.219**	**20.09**	**0.583**	**0.313**	**19.24**	**0.570**	**0.336**

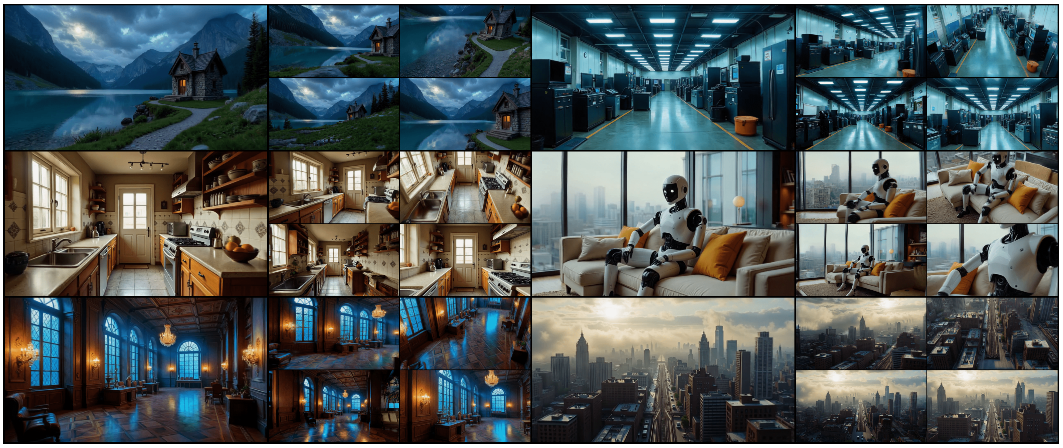


Figure 6: **Image-to-3DGS Generation.** We visualize five views from generated 3DGS scenes.

6 Experiments

6.1 Experimental Setup

Datasets. To train our model, we do not use any existing multi-view datasets; instead, we construct our own, which we call the *Lyra* dataset, relying solely on our video model to supervise the 3DGS decoder. The 3D reconstruction setup uses 59,031 images, while the 4D setup has 7,378 videos. All data are from diverse text prompts, spanning scenarios such as indoor and outdoor environments, humans, animals, and both realistic and imaginative content. We synthesize six camera trajectories for each image (3D) or video (4D), yielding 354,186 videos for 3D and 44,268 videos for 4D.

Baselines. We compare our method with the state-of-the-art approaches for generative single-image-to-3D reconstruction, i.e., ZeroNVS (Sargent et al., 2024), ViewCrafter (Yu et al., 2024b), Wonderland (Liang et al., 2025a), and Bolt3D (Szymanowicz et al., 2025b). Note that *no source code is available* to evaluate these methods on our out-of-distribution evaluation set, hence we mainly rely on reported quantitative comparisons from the papers.

Evaluation. We follow previous works to evaluate our model on the task of single-image to 3D using RealEstate10K (Zhou et al., 2018), DL3DV (Ling et al., 2024b), and Tanks and Temples (Knapitsch et al., 2017). We follow the evaluation protocol outlined in Wonderland (Liang et al., 2025a) and Bolt3D (Szymanowicz et al., 2025b). We evaluate the performance using standard reconstruction metrics, i.e., PSNR, SSIM, and LPIPS.

6.2 Main Results

Quantitative results. We show quantitative evaluations for single image-to-3D in Tab. 1. Our method outperforms previous works on all benchmarks across all metrics. Hence, the main improvements to further boost the quality lie in the development of stronger video generative models, as our reconstructions will directly benefit from them. We provide additional quantitative comparisons on the *Lyra* dataset in Appendix D.

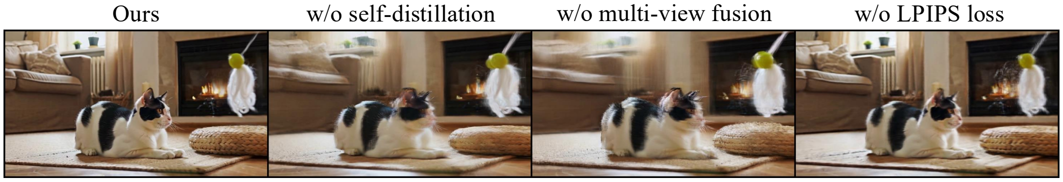


Figure 7: **Ablations.** We visualize ablation results by rendering the same extreme novel viewpoint after image-to-3DGS generation; depth visualizations and ablations are provided in Appendix D.4.

Qualitative results. We visualize novel view renderings from our generated 3DGS scenes in Fig. 1 and Fig. 6, but strongly recommend the reader to our supplementary webpage for video results. Our method generates high-quality novel view content for 3D/4D scenes in unseen regions while maintaining consistency with the input image/video. We visualize qualitative comparisons on the *Lyra* dataset in Appendix D.

6.3 Ablations

In Tab. 2, we motivate our design choices by ablating key components on out-of-distribution diverse prompts. We compare 3DGS renderings with videos generated by our camera-controlled video diffusion model.

Real data only. Instead of using self-distillation, we train a model using only real multi-view datasets, i.e., RealEstate10K (Zhou et al., 2018) and DL3DV (Ling et al., 2024b). These datasets are commonly used in previous works, but lead to limited generalizability to out-of-distribution scenes.

Self-distillation + real data. Perhaps surprisingly, joint training with self-distillation and real data does not improve over using only self-distillation without real data. This confirms that self-distillation is diverse and consistent enough to learn a reconstruction model.

No depth loss. Depth loss prevents flat geometry and even slightly improves image-based metrics. We visualize depth maps in Appendix D.4.

No opacity pruning. We keep all Gaussians instead of pruning them by opacity. Learning to prune most of the 3D Gaussians makes the output representation more compact and slightly improves visual quality. Rendering at $H = 704$, and $W = 1280$ takes 18ms with pruning vs. 30ms without pruning, a $1.67\times$ speed up.

Table 2: **Ablation study on *Lyra* dataset.**

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	24.77	0.837	0.224
Data			
real data only	19.08	0.659	0.413
self-distill. + real data	24.74	0.823	0.236
Loss			
w/o depth loss	24.31	0.811	0.247
w/o opacity pruning	24.55	0.820	0.237
w/o LPIPS loss	23.74	0.766	0.370
Architecture			
w/o multi-view fusion	17.73	0.632	0.446
w/o Mamba-2	24.58	0.818	0.241
w/o latent 3DGS		OOM	

No LPIPS loss. Adding the LPIPS loss improves the robustness to input inconsistencies and enhances the preservation of high-frequency details.

No multi-view fusion. A naive implementation would be to generate 3D Gaussians for each camera trajectory independently and then fuse the 3D Gaussians into one point cloud. However, we observe significant advantages by learning the multi-view fusion from different trajectories with our reconstruction blocks, where each token attends to the others.

No Mamba-2. We replace our joint Transformer/Mamba-2 with Transformer-only blocks and observe slightly lower quality. One forward pass with $V = 6$, $L = 121$, $H = 704$, and $W = 1280$ takes 3213ms with joint blocks in comparison to 20922ms with Transformer-only blocks, a $6.5 \times$ speedup.

No latent-based 3DGS. Previous works operating in the pixel space, such as BTimer (Liang et al., 2025b), are restricted to 12 input frames, while we can take up to 726 input frames. Consequently, operating the 3DGS decoder in the pixel space instead of the latent space leads to out-of-memory.

7 Conclusion

In this work, we propose *Lyra*, a novel 3D and 4D generation framework relying only on a single image or video input. Instead of collecting multi-view datasets, we introduce camera-controlled video diffusion models as teachers within a self-distillation framework for a student 3DGS decoder. Our 3DGS decoder operates in the latent space of the video model and directly reconstructs 3D Gaussians without any post-processing or optimization. This design enhances generalizability and coverage across diverse scenes, while our dynamic extension demonstrates the feasibility of 4D generation from monocular video. Currently, the scale and consistency of our generated scenes are bounded by the capacity of our camera-controlled video diffusion model. Hence, further improving the video diffusion model will enable large-scale, consistent scene synthesis. Moreover, investigating the adaptation of auto-regressive techniques (Chen et al., 2024a) into our framework for large-scale generation is an interesting direction for future work. Lastly, modelling motion and tracking information within the reconstruction network, as in concurrent work (Lin et al., 2025), could improve visual motion quality.

8 Ethics Statement

Like all generative AI technologies, there are risks of misuse of our method, including producing misleading or synthetic 3D/4D content. While such risks exist, the main goal of this research is to contribute to the fields of simulation, robotics, and embodied AI by providing scalable and controllable tools for data generation. In this sense, we expect our most significant contributions to be made through the following capabilities and insights.

- **The ability to enable realistic simulation for embodied agents**: We provide controllable and physically consistent camera motion within generated 3D and 4D environments to support training and evaluating agents that must perceive, navigate, and interact with complex scenes.
- **Improved scalability for data engines**: Our method removes the need for multi-view capture and expensive per-scene optimization, and helps produce high-fidelity, customizable scenes at scale, which is particularly suitable for research in robotics, reinforcement learning, and closed-loop simulation.
- **Better technical understanding**: As an academic study, this work contributes a foundational case study of how camera control interacts with spatiotemporal generative models, which will benefit research in computer vision, graphics, and embodied AI.

While we caution against deceptive or wrongful use of this technology—where misleading and/or unverifiable media could be created with realistic generative outputs—we believe that the adoption of safeguards (e.g., provenance tracking, dataset documentation, and good evaluation practices) will be useful in ensuring that generative models advance science and society in positive ways.

9 Reproducibility Statement

To fully reproduce our results and accelerate future research in this area, we release our training and inference code, model weights, and data for both 3D and 4D generation.

10 Acknowledgements

We thank Zian Wang for feedback on the draft.

References

Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Hao Tang, Gordon Wetzstein, Leonidas Guibas, Luc Van Gool, and Radu Timofte. 3d-aware video generation. In *TMLR*, 2023a.

Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon Wetzstein, Leonidas Guibas, and Andrea Tagliasacchi. CC3D: Layout-conditioned generation of compositional 3D scenes. In *Proc. ICCV*, 2023b.

Sherwin Bahmani, Xian Liu, Wang Yifan, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, Andrea Tagliasacchi, and David B. Lindell. Tc4d: Trajectory-conditioned text-to-4d generation. In *Proc. ECCV*, 2024a.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B. Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proc. CVPR*, 2024b.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. *Proc. CVPR*, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. *Proc. ICLR*, 2025b.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. *Proc. ICLR*, 2025b.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

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

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. *arXiv preprint arXiv:2504.14899*, 2025.

Yukang Cao, Liang Pan, Kai Han, Kwan-Yee K Wong, and Ziwei Liu. Avatargo: Zero-shot 4d human-object interaction generation and animation. *arXiv preprint arXiv:2410.07164*, 2024.

Zenghao Chai, Chen Tang, Yongkang Wong, and Mohan Kankanhalli. Star: Skeleton-aware text-based 4d avatar generation with in-network motion retargeting. *arXiv preprint arXiv:2406.04629*, 2024.

Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3D generative adversarial networks. In *Proc. CVPR*, 2022.

Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proc. ICCV*, 2023.

David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proc. CVPR*, 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Proc. NeurIPS*, 2024a.

Ce Chen, Shaoli Huang, Xuelin Chen, Guangyi Chen, Xiaoguang Han, Kun Zhang, and Mingming Gong. Ct4d: Consistent text-to-4d generation with animatable meshes. *arXiv preprint arXiv:2408.08342*, 2024b.

Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023a.

Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2Shape: Generating shapes from natural language by learning joint embeddings. In *Proc. ACCV*, 2018.

Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3D: Disentangling geometry and appearance for high-quality text-to-3D content creation. *arXiv preprint arXiv:2303.13873*, 2023b.

Soon Yau Cheong, Duygu Ceylan, Armin Mustafa, Andrew Gilbert, and Chun-Hao Paul Huang. Boosting camera motion control for video diffusion transformers. *arXiv preprint arXiv:2410.10802*, 2024.

Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024.

Team Cosmos. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W Taylor, and Joshua M Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proc. ICCV*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proc. ICLR*, 2021.

Qijun Feng, Zhen Xing, Zuxuan Wu, and Yu-Gang Jiang. FDGaussian: Fast Gaussian splatting from single image via geometric-aware diffusion model. *arXiv preprint arXiv:2403.10242*, 2024a.

Yutao Feng, Yintong Shang, Xiang Feng, Lei Lan, Shandian Zhe, Tianjia Shao, Hongzhi Wu, Kun Zhou, Hao Su, Chenfanfu Jiang, et al. Elastogen: 4d generative elastodynamics. *arXiv preprint arXiv:2405.15056*, 2024b.

Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. Gaussianflow: Splatting Gaussian dynamics for 4D content creation. *arXiv preprint arXiv:2403.12365*, 2024a.

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

William Gao, Noam Aigerman, Thibault Groueix, Vova Kim, and Rana Hanocka. TextDeformer: Geometry manipulation using text guidance. In *SIGGRAPH*, 2023.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. NerfDiff: Single-image view synthesis with NeRF-guided distillation from 3D-aware diffusion. In *Proc. ICML*, 2023.

Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. Diffusion as shader: 3d-aware video diffusion for versatile video generation control. In *SIGGRAPH*, 2025.

Junlin Han, Filippos Kokkinos, and Philip Torr. VFusion3D: Learning scalable 3D generative models from video diffusion models. *arXiv preprint arXiv:2403.12034*, 2024.

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*, 2024a.

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. *arXiv preprint arXiv:2503.10592*, 2025.

Xianglong He, Junyi Chen, Sida Peng, Di Huang, Yangguang Li, Xiaoshui Huang, Chun Yuan, Wanli Ouyang, and Tong He. GVGEN: Text-to-3D generation with volumetric representation. *arXiv preprint arXiv:2403.12957*, 2024b.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *Proc. ICCV*, 2023.

Lukas Höllein, Aljaž Božič, Norman Müller, David Novotny, Hung-Yu Tseng, Christian Richardt, Michael Zollhöfer, and Matthias Nießner. ViewDiff: 3D-consistent image generation with text-to-image models. In *Proc. CVPR*, 2024.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3D. In *Proc. ICLR*, 2024.

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

Ronghang Hu, Nikhila Ravi, Alexander C Berg, and Deepak Pathak. Worldsheet: Wrapping the world in a 3d sheet for view synthesis from a single image. In *Proc. ICCV*, 2021.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Teng Hu, Jiangning Zhang, Ran Yi, Yating Wang, Hongrui Huang, Jieyu Weng, Yabiao Wang, and Lizhuang Ma. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*, 2024.

Chun-Hao Paul Huang, Jae Shin Yoon, Hyeonho Jeong, Niloy Mitra, and Duygu Ceylan. Camera pose estimation emerging in video diffusion transformer, 2024a. URL `https://openreview.net/forum?id=1gf2LW7fOJ`.

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. In *NVIDIA Research Whitepapers*, 2025.

Tianyu Huang, Yihan Zeng, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physical properties of dynamic 3d gaussians with video diffusion priors. *arXiv preprint arXiv:2406.01476*, 2024b.

Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proc. CVPR*, 2022.

Nikolay Jetchev. ClipMatrix: Text-controlled creation of 3D textured meshes. *arXiv preprint arXiv:2109.12922*, 2021.

Lihan Jiang, Yucheng Mao, Linning Xu, Tao Lu, Kerui Ren, Yichen Jin, Xudong Xu, Mulin Yu, Jiangmiao Pang, Feng Zhao, et al. Anysplat: Feed-forward 3d gaussian splatting from unconstrained views. *arXiv preprint arXiv:2505.23716*, 2025a.

Lutao Jiang and Lin Wang. Brightdreamer: Generic 3D Gaussian generative framework for fast text-to-3D synthesis. *arXiv preprint arXiv:2403.11273*, 2024.

Yanqin Jiang, Chaohui Yu, Chenjie Cao, Fan Wang, Weiming Hu, and Jin Gao. Animate3d: Animating any 3d model with multi-view video diffusion. *arXiv preprint arXiv:2407.11398*, 2024a.

Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. In *Proc. ICLR*, 2024b.

Zeren Jiang, Chuanxia Zheng, Iro Laina, Diane Larlus, and Andrea Vedaldi. Geo4d: Leveraging video generators for geometric 4d scene reconstruction. *arXiv preprint arXiv:2504.07961*, 2025b.

Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *Proc. CVPR*, 2025.

Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. *arXiv preprint arXiv:2503.19907*, 2025.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. Spad: Spatially aware multi-view diffusers. In *Proc. CVPR*, 2024.

Yash Kant, Ethan Weber, Jin Kyu Kim, Rawal Khirodkar, Su Zhaoen, Julieta Martinez, Igor Gilitschenski, Shunsuke Saito, and Timur Bagautdinov. Pippo: High-resolution multi-view humans from a single image. In *Proc. CVPR*, 2025.

Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. In *Proc. ICLR*, 2024.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM TOG*, 2023.

Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, and Sanja Fidler. NeuralField-LDM: Scene generation with hierarchical latent diffusion models. In *Proc. CVPR*, 2023.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Mert Kiray, Paul Uhlenbruck, Nassir Navab, and Benjamin Busam. Promptvfx: Text-driven fields for open-world 3d gaussian animation. *arXiv preprint arXiv:2506.01091*, 2025.

Tobias Kirschstein, Javier Romero, Artem Sevastopolsky, Matthias Nießner, and Shunsuke Saito. Avat3r: Large animatable gaussian reconstruction model for high-fidelity 3d head avatars. *arXiv preprint arXiv:2502.20220*, 2025.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017.

Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: Consistent multi-video generation with camera control. In *Proc. NeurIPS*, 2024.

Kyungmin Lee, Kihyuk Sohn, and Jinwoo Shin. DreamFlow: High-quality text-to-3D generation by approximating probability flow. In *Proc. ICLR*, 2024a.

Yao-Chih Lee, Yi-Ting Chen, Andrew Wang, Ting-Hsuan Liao, Brandon Y Feng, and Jia-Bin Huang. Vividdream: Generating 3d scene with ambient dynamics. *arXiv preprint arXiv:2405.20334*, 2024b.

Guojun Lei, Chi Wang, Hong Li, Rong Zhang, Yikai Wang, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. *arXiv preprint arXiv:2411.10836*, 2024.

Guojun Lei, Chi Wang, Rong Zhang, Yikai Wang, Hong Li, and Weiwei Xu. Animateanything: Consistent and controllable animation for video generation. In *Proc. CVPR*, 2025.

Bing Li, Cheng Zheng, Wenxuan Zhu, Jinjie Mai, Biao Zhang, Peter Wonka, and Bernard Ghanem. Vivid-zoo: Multi-view video generation with diffusion model. *arXiv preprint arXiv:2406.08659*, 2024a.

Jiahao Li, Hao Tan, Kai Zhang, Zexiang Xu, Fujun Luan, Yinghao Xu, Yicong Hong, Kalyan Sunkavalli, Greg Shakhnarovich, and Sai Bi. Instant3D: Fast text-to-3D with sparse-view generation and large reconstruction model. In *Proc. ICLR*, 2024b.

Jinwei Li, Huan-ang Gao, Wenyi Li, Haohan Chi, Chenyu Liu, Chenxi Du, Yiqian Liu, Mingju Gao, Guiyu Zhang, Zongzheng Zhang, et al. Fb-4d: Spatial-temporal coherent dynamic 3d content generation with feature banks. *arXiv preprint arXiv:2503.20784*, 2025a.

Longfei Li, Zhiwen Fan, Wenyan Cong, Xinhang Liu, Yuyang Yin, Matt Foutter, Panwang Pan, Chenyu You, Yue Wang, Zhangyang Wang, et al. Martian world models: Controllable video synthesis with physically accurate 3d reconstructions. *arXiv preprint arXiv:2507.07978*, 2025b.

Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, and Zhiwen Fan. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024c.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025c.

Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proc. CVPR*, 2022.

Zhiqi Li, Yiming Chen, and Peidong Liu. Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. *arXiv preprint arXiv:2410.06756*, 2024d.

Zhiqi Li, Yiming Chen, Lingzhe Zhao, and Peidong Liu. Controllable text-to-3D generation via surface-aligned Gaussian splatting. *arXiv preprint arXiv:2403.09981*, 2024e.

Hanwen Liang, Yuyang Yin, Dejia Xu, Hanxue Liang, Zhangyang Wang, Konstantinos N Plataniotis, Yao Zhao, and Yunchao Wei. Diffusion4d: Fast spatial-temporal consistent 4d generation via video diffusion models. *arXiv preprint arXiv:2405.16645*, 2024.

Hanwen Liang, Junli Cao, Vidit Goel, Guocheng Qian, Sergei Korolev, Demetri Terzopoulos, Konstantinos N Plataniotis, Sergey Tulyakov, and Jian Ren. Wonderland: Navigating 3d scenes from a single image. *Proc. CVPR*, 2025a.

Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, and Jiahui Huang. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. *Proc. NeurIPS*, 2025b.

Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3D generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3D: High-resolution text-to-3D content creation. In *Proc. CVPR*, 2023a.

Chenguo Lin, Yuchen Lin, Panwang Pan, Yifan Yu, Honglei Yan, Katerina Fragkiadaki, and Yadong Mu. Movies: Motion-aware 4d dynamic view synthesis in one second. *arXiv preprint arXiv:2507.10065*, 2025.

Jiajing Lin, Zhenzhong Wang, Yongjie Hou, Yuzhou Tang, and Min Jiang. Phy124: Fast physics-driven 4d content generation from a single image. *arXiv preprint arXiv:2409.07179*, 2024.

Yukang Lin, Haonan Han, Chaoqun Gong, Zunnan Xu, Yachao Zhang, and Xiu Li. Consistent123: One image to highly consistent 3D asset using case-aware diffusion priors. In *arXiv preprint arXiv:2309.17261*, 2023b.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proc. CVPR*, 2024a.

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 *Proc. CVPR*, 2024b.

Pengyang Ling, Jiazi Bu, Pan Zhang, Xiaoyi Dong, Yuhang Zang, Tong Wu, Huaian Chen, Jiaqi Wang, and Yi Jin. Motionclone: Training-free motion cloning for controllable video generation. *arXiv preprint arXiv:2406.05338*, 2024c.

Lijuan Liu, Wenfa Li, Dongbo Zhang, Shuo Wang, and Shaohui Jiao. Idcnet: Guided video diffusion for metric-consistent rgbd scene generation with precise camera control. *arXiv preprint arXiv:2508.04147*, 2025.

Pengkun Liu, Yikai Wang, Fuchun Sun, Jiafang Li, Hang Xiao, Hongxiang Xue, and Xinzhou Wang. Isotropic3D: Image-to-3D generation based on a single clip embedding. *arXiv preprint arXiv:2403.10395*, 2024a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *Proc. ICCV*, 2023.

Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. HumanGaussian: Text-driven 3D human generation with Gaussian splatting. In *Proc. CVPR*, 2024b.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *Proc. ICLR*, 2024c.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3D: Single image to 3D using cross-domain diffusion. In *Proc. CVPR*, 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. *arXiv preprint arXiv:2412.03934*, 2024.

Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. Follow-your-creation: Empowering 4d creation through video inpainting. *arXiv preprint arXiv:2506.04590*, 2025.

Jinjie Mai, Wenxuan Zhu, Haozhe Liu, Bing Li, Cheng Zheng, Jürgen Schmidhuber, and Bernard Ghanem. Can video diffusion model reconstruct 4d geometry? *arXiv preprint arXiv:2503.21082*, 2025.

Qiaowei Miao, Yawei Luo, and Yi Yang. Pla4d: Pixel-level alignments for text-to-4d gaussian splatting. *arXiv preprint arXiv:2405.19957*, 2024.

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. In *Proc. ECCV*, 2020.

Nicolas Moenne-Loccoz, Ashkan Mirzaei, Or Perel, Riccardo de Lutio, Janick Martinez Esturo, Gavriel State, Sanja Fidler, Nicholas Sharp, and Zan Gojcic. 3d gaussian ray tracing: Fast tracing of particle scenes. *ACM Transactions on Graphics and SIGGRAPH Asia*, 2024.

OpenAI. Chatgpt (gpt-5), 2025. Large language model. Available at `https://chat.openai.com/`.

Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-resolution 3D-consistent image and geometry generation. In *Proc. CVPR*, 2022.

Zijie Pan, Zeyu Yang, Xiatian Zhu, and Li Zhang. Fast dynamic 3d object generation from a single-view video. *arXiv preprint arXiv:2401.08742*, 2024.

Jangho Park, Taesung Kwon, and Jong Chul Ye. Zero4d: Training-free 4d video generation from single video using off-the-shelf video diffusion. *arXiv preprint arXiv:2503.22622*, 2025.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *Proc. ICLR*, 2023.

Guocheng Qian, Junli Cao, Aliaksandr Siarohin, Yash Kant, Chaoyang Wang, Michael Vasilkovsky, Hsin-Ying Lee, Yuwei Fang, Ivan Skorokhodov, Peiye Zhuang, et al. Atom: Amortized text-to-mesh using 2d diffusion. *arXiv preprint arXiv:2402.00867*, 2024a.

Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, et al. Magic123: One image to high-quality 3D object generation using both 2D and 3D diffusion priors. In *Proc. ICLR*, 2024b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. DreamGaussian4D: Generative 4D Gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung Wook Kim, et al. L4gm: Large 4d gaussian reconstruction model. In *Proc. NeurIPS*, 2024a.

Xuanchi Ren, Jiahui Huang, Xiaohui Zeng, Ken Museth, Sanja Fidler, and Francis Williams. Xcube: Large-scale 3d generative modeling using sparse voxel hierarchies. In *Proc. CVPR*, 2024b.

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. *Proc. NeurIPS*, 2024c.

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. *Proc. CVPR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022.

Aditya Sanghi, Hang Chu, Joseph G Lambourne, Ye Wang, Chin-Yi Cheng, Marco Fumero, and Kamal Rahimi Malekshan. CLIP-Forge: Towards zero-shot text-to-shape generation. In *Proc. CVPR*, 2022.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single image. In *Proc. CVPR*, 2024.

Manuel-Andreas Schneider, Lukas Höllein, and Matthias Nießner. Worldexplorer: Towards generating fully navigable 3d scenes. *arXiv preprint arXiv:2506.01799*, 2025.

Katja Schwarz, Axel Sauer, Michael Niemeyer, Yiyi Liao, and Andreas Geiger. VoxGRAF: Fast 3D-aware image synthesis with sparse voxel grids. In *Proc. NeurIPS*, 2022.

Katja Schwarz, Norman Mueller, and Peter Kontschieder. Generative gaussian splatting: Generating 3d scenes with video diffusion priors. *arXiv preprint arXiv:2503.13272*, 2025.

Junyoung Seo, Jisang Han, Jaewoo Jung, Siyoon Jin, Joungbin Lee, Takuya Narihira, Kazumi Fukuda, Takashi Shibuya, Donghoon Ahn, Shoukang Hu, et al. Vid-camedit: Video camera trajectory editing with generative rendering from estimated geometry. *arXiv preprint arXiv:2506.13697*, 2025.

Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3D generation. In *Proc. ICLR*, 2024.

Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. Realmdreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199*, 2024.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. In *Proc. ICML*, 2023.

Jingxiang Sun, Bo Zhang, Ruizhi Shao, Lizhen Wang, Wen Liu, Zhenda Xie, and Yebin Liu. DreamCraft3D: Hierarchical 3D generation with bootstrapped diffusion prior. In *Proc. ICLR*, 2024a.

Keqiang Sun, Dor Litvak, Yunzhi Zhang, Hongsheng Li, Jiajun Wu, and Shangzhe Wu. Ponymation: Learning articulated 3d animal motions from unlabeled online videos. In *Proc. ECCV*, 2024b.

Qi Sun, Zhiyang Guo, Ziyu Wan, Jing Nathan Yan, Shengming Yin, Wengang Zhou, Jing Liao, and Houqiang Li. Eg4d: Explicit generation of 4d object without score distillation. *arXiv preprint arXiv:2405.18132*, 2024c.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024d.

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Viewset diffusion:(0-) image-conditioned 3d generative models from 2d data. In *Proc. ICCV*, 2023.

Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3D reconstruction. In *Proc. CVPR*, 2024.

Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, João F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *Proc. 3DV*, 2025a.

Stanislaw Szymanowicz, Jason Y Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T Barron, and Philipp Henzler. Bolt3d: Generating 3d scenes in seconds. *arXiv preprint arXiv:2503.14445*, 2025b.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. In *Proc. ECCV*, 2024a.

Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. LGM: Large multi-view gaussian model for high-resolution 3d content creation. *Proc. ECCV*, 2024b.

Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3D: High-fidelity 3D creation from a single image with diffusion prior. *arXiv preprint arXiv:2303.14184*, 2023a.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *Proc. NeurIPS*, 2023b.

Zhenggang Tang, Peiye Zhuang, Chaoyang Wang, Aliaksandr Siarohin, Yash Kant, Alexander Schwing, Sergey Tulyakov, and Hsin-Ying Lee. Pixel-aligned multi-view generation with depth guided decoder. *arXiv preprint arXiv:2408.14016*, 2024c.

Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Josh Tenenbaum, Frédo Durand, Bill Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *Proc. NeurIPS*, 2023.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. Triposr: Fast 3D object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

Lukas Uzolas, Elmar Eisemann, and Petr Kellnhofer. Motiondreamer: Zero-shot 3d mesh animation from video diffusion models. *arXiv preprint arXiv:2405.20155*, 2024.

Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *Proc. ECCV*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, 2017.

Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Ziyu Wan, Despoina Paschalidou, Ian Huang, Hongyu Liu, Bokui Shen, Xiaoyu Xiang, Jing Liao, and Leonidas Guibas. CAD: Photorealistic 3D generation via adversarial distillation. In *Proc. CVPR*, 2024.

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-NeRF: Text-and-image driven manipulation of neural radiance fields. In *Proc. CVPR*, 2022.

Chaoyang Wang, Peiye Zhuang, Tuan Duc Ngo, Willi Menapace, Aliaksandr Siarohin, Michael Vasilkovsky, Ivan Skorokhodov, Sergey Tulyakov, Peter Wonka, and Hsin-Ying Lee. 4real-video: Learning generalizable photo-realistic 4d video diffusion. *Proc. CVPR*, 2025a.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score Jacobian chaining: Lifting pretrained 2d diffusion models for 3D generation. In *Proc. CVPR*, 2023a.

Peng Wang and Yichun Shi. Imagedream: Image-prompt multi-view diffusion for 3d generation. *arXiv preprint arXiv:2312.02201*, 2023.

Qinghe Wang, Yawen Luo, Xiaoyu Shi, Xu Jia, Huchuan Lu, Tianfan Xue, Xintao Wang, Pengfei Wan, Di Zhang, and Kun Gai. Cinemaster: A 3d-aware and controllable framework for cinematic text-to-video generation. In *SIGGRAPH*, 2025b.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision, 2024a. URL `https://arxiv.org/abs/2410.19115`.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proc. CVPR*, 2024b.

Xi Wang, Robin Courant, Marc Christie, and Vicky Kalogeiton. Akira: Augmentation kit on rays for optical video generation. In *Proc. CVPR*, 2025c.

Xiaodong Wang, Zhirong Wu, and Peixi Peng. Longdwm: Cross-granularity distillation for building a long-term driving world model. *arXiv preprint arXiv:2506.01546*, 2025d.

Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels. *arXiv preprint arXiv:2405.16822*, 2024c.

Yikai Wang, Guangce Liu, Xinzhou Wang, Zilong Chen, Jiafang Li, Xin Liang, Fuchun Sun, and Jun Zhu. Video4dgen: Enhancing video and 4d generation through mutual optimization. *arXiv preprint arXiv:2504.04153*, 2025e.

Yizhi Wang, Mingrui Zhao, Ali Mahdavi-Amiri, and Hao Zhang. Act-r: Adaptive camera trajectories for single view 3d reconstruction. *arXiv preprint arXiv:2505.08239*, 2025f.

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*, 2024d.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolific-Dreamer: High-fidelity and diverse text-to-3D generation with variational score distillation. In *Proc. NeurIPS*, 2023b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Yin Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH*, 2024e.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *Proc. ICLR*, 2023.

Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024.

Qi Wu, Janick Martinez Esturo, Ashkan Mirzaei, Nicolas Moenne-Loccoz, and Zan Gojcic. 3dgut: Enabling distorted cameras and secondary rays in gaussian splatting. In *Proc. CVPR*, 2025a.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. *Proc. CVPR*, 2025b.

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*, 2025c.

Zijie Wu, Chaohui Yu, Yanqin Jiang, Chenjie Cao, Fan Wang, and Xiang Bai. Sc4d: Sparse-controlled video-to-4d generation and motion transfer. *arXiv preprint arXiv:2404.03736*, 2024.

Zeqi Xiao, Yifan Zhou, Shuai Yang, and Xingang Pan. Video diffusion models are training-free motion interpreter and controller. *arXiv preprint arXiv:2405.14864*, 2024.

Kevin Xie, Jonathan Lorraine, Tianshi Cao, Jun Gao, James Lucas, Antonio Torralba, Sanja Fidler, and Xiaohui Zeng. LATTE3D: Large-scale amortized text-to-enhanced3D synthesis. In *Proc. ECCV*, 2024a.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. *arXiv preprint arXiv:2407.17470*, 2024b.

Jinbo Xing, Long Mai, Cusuh Ham, Jiahui Huang, Aniruddha Mahapatra, Chi-Wing Fu, Tien-Tsin Wong, and Feng Liu. Motioncanvas: Cinematic shot design with controllable image-to-video generation. In *SIGGRAPH*, 2025.

Dejia Xu, Yifan Jiang, Chen Huang, Liangchen Song, Thorsten Gernoth, Liangliang Cao, Zhangyang Wang, and Hao Tang. Cavia: Camera-controllable multi-view video diffusion with view-integrated attention. *arXiv preprint arXiv:2410.10774*, 2024a.

Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024b.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024c.

Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. GRM: Large Gaussian reconstruction model for efficient 3D reconstruction and generation. In *Proc. ECCV*, 2024d.

Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, et al. DMV3D: Denoising multi-view diffusion using 3D large reconstruction model. In *Proc. ICLR*, 2024e.

Zhen Xu, Zhengqin Li, Zhao Dong, Xiaowei Zhou, Richard Newcombe, and Zhaoyang Lv. 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. *arXiv preprint arXiv:2506.08015*, 2025.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Wenqing Zhang, Song Bai, Jiashi Feng, and Mike Zheng Shou. Pv3d: A 3d generative model for portrait video generation. In *Proc. ICLR*, 2023.

Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, et al. Streetcrafter: Street view synthesis with controllable video diffusion models. In *Proc. CVPR*, 2025.

Qitong Yang, Mingtao Feng, Zijie Wu, Shijie Sun, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Beyond skeletons: Integrative latent mapping for coherent 4d sequence generation. *arXiv preprint arXiv:2403.13238*, 2024a.

Zeyu Yang, Zijie Pan, Chun Gu, and Li Zhang. Diffusion²: Dynamic 3d content generation via score composition of orthogonal diffusion models. *arXiv preprint 2404.02148*, 2024b.

Zhongqi Yang, Wenhang Ge, Yuqi Li, Jiaqi Chen, Haoyuan Li, Mengyin An, Fei Kang, Hua Xue, Baixin Xu, Yuyang Yin, et al. Matrix-3d: Omnidirectional explorable 3d world generation. *arXiv preprint arXiv:2508.08086*, 2025.

Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. DreamReward: Text-to-3D generation with human preference. *arXiv preprint arXiv:2403.14613*, 2024.

Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *JMLR*, 2025.

Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023.

Paul Yoo, Jiaxian Guo, Yutaka Matsuo, and Shixiang Shane Gu. DreamSparse: Escaping from Plato's cave with 2D diffusion model given sparse views. In *arXiv preprint arXiv:2306.03414*, 2023.

Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. In *Proc. NeurIPS*, 2024a.

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

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *Proc. ICCV*, 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*, 2024b.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3D with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023.

Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *Proc. CVPR*, 2025.

Yu-Jie Yuan, Leif Kobbelt, Jiwen Liu, Yuan Zhang, Pengfei Wan, Yu-Kun Lai, and Lin Gao. 4dynamic: Text-to-4d generation with hybrid priors. *arXiv preprint arXiv:2407.12684*, 2024.

Bohan Zeng, Ling Yang, Siyu Li, Jiaming Liu, Zixiang Zhang, Juanxi Tian, Kaixin Zhu, Yongzhen Guo, Fu-Yun Wang, Minkai Xu, et al. Trans4d: Realistic geometry-aware transition for compositional text-to-4d synthesis. *arXiv preprint arXiv:2410.07155*, 2024a.

Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024b.

Bowen Zhang, Tianyu Yang, Yu Li, Lei Zhang, and Xi Zhao. Compress3D: a compressed latent space for 3D generation from a single image. *arXiv preprint arXiv:2403.13524*, 2024a.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024b.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. 4diffusion: Multi-view video diffusion model for 4d generation. *arXiv preprint arXiv:2405.20674*, 2024c.

Hao Zhang, Di Chang, Fang Li, Mohammad Soleymani, and Narendra Ahuja. Magicpose4d: Crafting articulated models with appearance and motion control. *arXiv preprint arXiv:2405.14017*, 2024d.

Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *Proc. ECCV*, 2024e.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. CVPR*, 2018.

Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. Spatialcrafter: Unleashing the imagination of video diffusion models for scene reconstruction from limited observations. *arXiv preprint arXiv:2505.11992*, 2025a.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *Proc. ECCV*, 2024f.

Yisu Zhang, Chenjie Cao, Chaohui Yu, and Jianke Zhu. Lion-lora: Rethinking lora fusion to unify controllable spatial and temporal generation for video diffusion. *arXiv preprint arXiv:2507.05678*, 2025b.

Zhichao Zhang, Hui Chen, Jinsheng Deng, Xiaoqing Yin, Xingshen Song, and Ming Xu. Motion4d: A decoupled pipeline for enhanced text-to-4d generation with optimized motion patterns. *SSRN*, 2024g.

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024a.

Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. In *Proc. CVPR*, 2024b.

Junwei Zhou, Xueting Li, Lu Qi, and Ming-Hsuan Yang. Coco4d: Comprehensive and complex 4d scene generation. *arXiv preprint arXiv:2506.19798*, 2025.

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

Hanxin Zhu, Tianyu He, Anni Tang, Junliang Guo, Zhibo Chen, and Jiang Bian. Compositional 3d-aware video generation with llm director. *arXiv preprint arXiv:2409.00558*, 2024.

Hanxin Zhu, Tianyu He, Xiqian Yu, Junliang Guo, Zhibo Chen, and Jiang Bian. Ar4d: Autoregressive 4d generation from monocular videos. *arXiv preprint arXiv:2501.01722*, 2025.

Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. *arXiv preprint arXiv:2410.12781*, 2024.

A Supplementary Webpage

We provide video results on our supplementary webpage: `https://research.nvidia.com/labs/toronto-ai/lyra`.

B Additional Details of Video Diffusion Model

This section provides additional information on the GEN3C (Ren et al., 2025) video diffusion model backbone introduced in Sec. 3.1 and how we further improve its 3D consistency.

Conservative mask refinement. A key limitation of GEN3C's forward warping is background leakage, where occluded object parts in the source view (e.g., lion body in Fig. 8) become visible in novel viewpoints but are not properly indicated in the disocclusion mask. This leads to incomplete foreground completion by the video diffusion model, as shown in Fig. 8 (second and fourth columns).

To address this, we adopt a conservative strategy: since the geometry of occluded regions in $\mathbf{P}^{t,v}$ is unknown, we assume all such areas are occupied and derive refined disocclusion masks accordingly. This allows the video model to reason about potentially visible foreground regions that were missed by standard point-based rendering.

Our approach constructs a triangular mesh from camera-space points $\mathbf{P}_c^{t,v} = \mathbf{C}^t \mathbf{P}^{t,v}$ by connecting spatially adjacent pixels, similar to Hu et al. (2021):

$$\mathcal{M} = \bigcup_{(u,v)} \left\{ \triangle\left(\mathbf{p}_{u,v}, \mathbf{p}_{u+1,v}, \mathbf{p}_{u,v+1}\right), \triangle\left(\mathbf{p}_{u+1,v}, \mathbf{p}_{u+1,v+1}, \mathbf{p}_{u,v+1}\right) \right\} \tag{2}$$

where $\mathbf{p}_{u,v}$ denotes the 3D point at pixel (u, v) in $\mathbf{P}_c^{t,v}$. This surface mesh serves as a boundary between visible and invisible regions. For each rendered pixel, we compare the standard point-based depth $\mathbf{D}^{t,v}$ with the mesh-interpolated depth $\mathbf{D}_{\mathcal{M}}^{t,v}$ obtained via ray-surface intersection:

$$\mathbf{M}^{t,v}(u,v) = \begin{cases} 0 & \text{if } \mathbf{D}_{\mathcal{M}}^{t,v}(u,v) < \mathbf{D}^{t,v}(u,v) - \epsilon \\ \mathbf{M}_{\text{orig}}^{t,v}(u,v) & \text{otherwise} \end{cases}, \tag{3}$$

where $\mathbf{M}_{\text{orig}}^{t,v}$ denotes the original disocclusion mask from forward warping and ϵ is a small tolerance factor. When the mesh surface is closer than the original points, we conservatively mask out these regions as potential foreground disocclusions. This produces more reliable guidance for video generation, especially under large viewpoint changes, as shown in Fig. 8 (third and fifth columns).

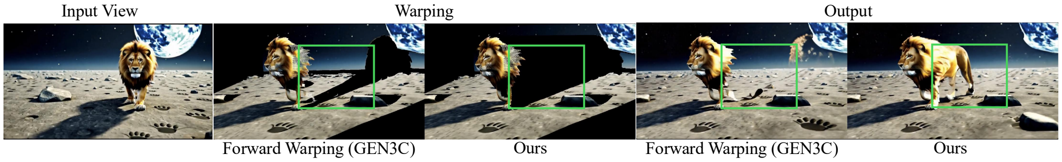


Figure 8: **Rendering function comparison.** Our model builds upon GEN3C (Ren et al., 2025) that uses forward warped images as camera control conditioning. We visualize forward warping (left) vs. our conservative rendering function (right). The highlighted region is incorrectly filled with background pixels in forward warping but properly masked in ours, enabling correct completion.

C Model Details

Progressive training. Directly training the 3DGS decoder on 6 trajectories of 121 frames at 704 × 1280 is expensive, hence we train our network progressively, as shown in Tab. 3. Furthermore, we use the final static pre-trained model to initialize our dynamic model. The total training time takes 6 days with 8 NVIDIA A100 80 GB GPUs. We use gsplat (Ye et al., 2025) as the 3DGS implementation.

Inference. One forward pass through the video diffusion model takes 12 minutes. As video models become faster, especially with progress in distillation into few-step models, this number is expected to decrease drastically with orthogonal improvements by the video generation community. Our 3DGS decoder takes only 3 seconds to generate 3D Gaussians from video latents. Lastly, we can render the 3D Gaussians in real-time after generation from novel viewpoints, where rendering a 704×1280 frame takes 18 ms.

Efficiency and number of input trajectories. Our approach is trained to handle a variable number of input camera trajectories. During training, we randomly vary the number of trajectories between 1 and 6, enabling the model to generalize to any number of trajectories within this range at inference time. While we primarily use 6 trajectories in our experiments to maximize view coverage, the model can operate with as few as a single trajectory.

Importantly, using fewer trajectories allows avoiding any computational overhead beyond standard camera-controlled video generation. The 3DGS decoder requires only 1 second for a single trajectory and 3 seconds for six trajectories to produce 3D Gaussians, which is comparable to decoding RGB frames in 2D video generation models. Using additional trajectories therefore provides a controllable trade-off between view coverage and inference speed, without affecting reconstruction quality when fewer trajectories are used.

Dataset generation cost. The training dataset is generated fully automatically, without manual screening or automated filtering, making the approach highly scalable. Dataset generation is dominated by the cost of the underlying video diffusion model, which currently takes 12 minutes per trajectory on a single NVIDIA A100 GPU. As a result, generating one training sample with 6 trajectories takes 72 minutes.

For the 3D dataset, we generate $59{,}031$ scenes with 6 camera trajectories each, corresponding to approximately $71{,}000$ GPU hours. For the 4D dataset, we generate $7{,}378$ scenes with 12 trajectories each, corresponding to approximately $18{,}000$ GPU hours. This data generation process is performed once, offline, after which any number of models can be trained on the resulting dataset.

Collecting real-world data with comparable diversity is practically infeasible, especially for the 4D case, which would require multiple cameras synchronized in time. We fully open-source the generated datasets to facilitate future research. As video generation models continue to become faster, the cost of dataset generation is expected to decrease substantially. In contrast, training the 3DGS decoder itself is relatively inexpensive, requiring 6 days on 8 NVIDIA A100 GPUs, whereas related approaches often require multi-node training.

Subsampling Gaussians. Generating pixel-aligned 3D Gaussians scales linearly with each spatial and temporal resolution. Our generated scenes represent 726 frames at 704×1280 resolution, leading to 654,213,120 per-pixel Gaussians. Hence, we only generate one 3D Gaussian per 8×8 spatial neighborhood, reducing the total number of 3D Gaussians by a factor of 64, i.e., 10,222,080. Moreover, our opacity-based pruning further reduces the number to 2,044,416, making the output more compact.

Joint Transformer-Mamba blocks. We use the reconstruction block design introduced in Long-LRM (Ziwen et al., 2024), i.e., one block consists of one Transformer layer followed by seven Mamba-2 layers. We repeat the block twice to get 16 layers with 512 hidden dimensions. Note that our network is significantly smaller than the one used in Long-LRM (Ziwen et al., 2024) or Wonderland (Liang et al., 2025a), which use 24 layers with 1024 hidden dimensions. While pure Transformers are typically more powerful, they are significantly slower in training and inference than Mamba2. Since we trained both configurations within the same training budget, we hypothesize that transformer-only blocks would need considerably more training iterations to fully catch up in rendering quality.

Camera encoding. We use Plücker embeddings for camera encoding, following prior works (Kant et al., 2024; Zhang et al., 2024e). The embeddings are computed at full spatial and temporal resolution and then projected into the video latent space using the pre-trained RGB encoder. Since the RGB encoder expects 3 input channels but Plücker embeddings are 6-dimensional, we separately encode the 3-dimensional ray directions and the 3-dimensional cross product of ray directions and origins. The resulting latent encodings are then concatenated along the channel dimension.

Time encoding. Similarly, to make the 1-dimensional time input compatible with the 3-channel input expected by the RGB encoder, we concatenate a 2-dimensional sinusoidal embedding to the original time value along the channel dimension. We then replicate these values across the spatial dimensions

Table 3: **Progressive training setup.**

Stage	$H \times W$	L	V	S	B	Steps
Static						
1	176×320	17	1	17	4	10k
2	176×320	49	1	49	4	2.5k
3	352×640	49	1	49	2	2.5k
4	704×1280	49	1	49	1	2.5k
5	704×1280	121	1	9	1	57.5k
6	704×1280	121	1–6	9	1	7k
Dynamic						
7	704×1280	121	6	12	1	10k

before encoding the time into the video latent space. This approach allows us to independently encode both the source and target times into the compressed latent representation.

Dynamic data augmentation. We visualize our dynamic data augmentation procedure for two example camera trajectories in Fig. 9. The procedure creates pairs of supervision views that cover the scene from the same motion state but different extreme viewpoints.

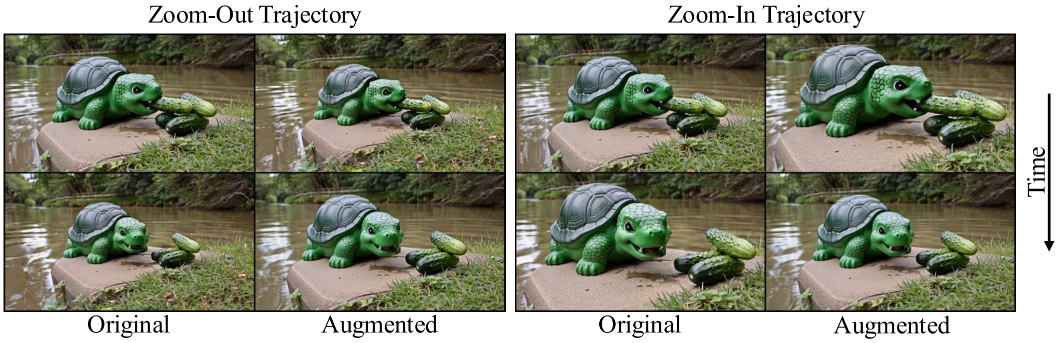


Figure 9: **Dynamic data augmentation videos.** We augment the supervision data with a motion-reversed video, ensuring that each timestep is observed from the full spatial coverage, thereby preventing low opacity artifacts in the early timesteps. We show two example trajectories, i.e., zoom-out and zoom-in, and visualize their corresponding augmented videos. The augmented videos are flipped in their camera motion.

Multi-view consistency. All camera trajectories are generated independently, but are rendered from a shared underlying 3D/4D cache. In our framework, the video generation backbone (GEN3C) models scene memory through this cache, enabling camera-controlled video synthesis with implicit scene consistency across trajectories.

While this shared cache already provides a degree of multi-view coherence, further improvements are possible by integrating more advanced memory mechanisms developed for long-form video generation. Recent approaches such as Context as Memory (Yu et al., 2025) explicitly retrieve and reuse scene information over extended temporal horizons, and could further strengthen multi-view consistency in a multi-trajectory setting. We leave the integration of such memory mechanisms for future work.

D Additional Evaluation

D.1 Additional Baseline details

We additionally compare our model against BTimer (Liang et al., 2025b), a recent approach for joint 3D and 4D reconstruction from posed images or videos. We reached out to the authors of BTimer to conduct experiments for us using the same scenes as *Lyra*. Since BTimer is purely regression-based, we integrate it with our GEN3C (Ren et al., 2025) video diffusion backbone. Specifically, we use the

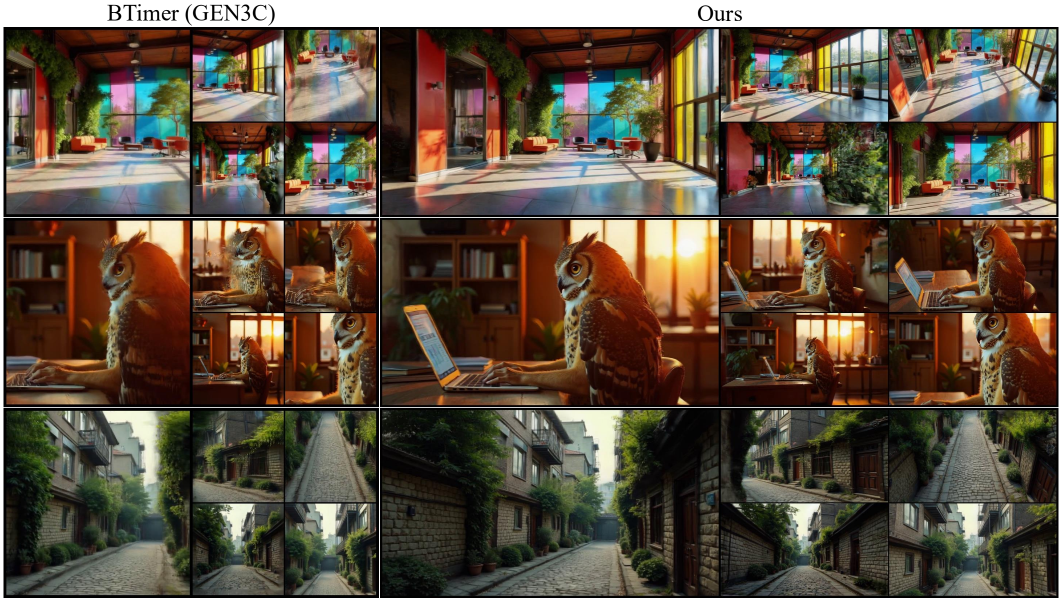


Figure 10: **Image-to-3DGS Comparison.** We compare our method with BTimer (GEN3C) and visualize five views from generated scenes. We observe significantly fewer artifacts and higher fidelity.

Table 4: **Comparison with pixel-space Long-LRM (GEN3C) baseline.** Quantitative results on RE10K, DL3DV, and Tanks&Temples datasets.

Dataset	Method	PSNR ↑	SSIM ↑	LPIPS ↓
RE10K	Long-LRM (GEN3C)	18.45	0.653	0.378
	Ours	**21.79**	**0.752**	**0.219**
DL3DV	Long-LRM (GEN3C)	17.83	0.499	0.439
	Ours	**20.09**	**0.583**	**0.313**
T&T	Long-LRM (GEN3C)	16.56	0.491	0.482
	Ours	**19.24**	**0.570**	**0.336**

same camera trajectories as in our model and provide RGB-decoded videos from GEN3C as input to BTimer. BTimer operates in pixel space and requires 12 input images. However, as discussed in the main paper, pixel-space models such as BTimer run out of memory when directly using the 726 high-resolution frames generated by GEN3C. To address this, we uniformly subsample GEN3C frames. For static evaluations, we select 12 frames that evenly cover the available viewpoints. For dynamic evaluations, we sample 11 frames to span the viewpoint range and additionally include the bullet-time frame corresponding to the motion being reconstructed. This ensures that the target motion state is always present as a reference. We denote this integrated baseline as BTimer (GEN3C), i.e., BTimer combined with GEN3C.

D.2 Additional Static 3D Evaluation

Long-LRM + GEN3C We compare against a pixel-space baseline that combines GEN3C with Long-LRM (Ziwen et al., 2024). In this setting, GEN3C serves as the video generation backbone, while Long-LRM performs 3D reconstruction directly in pixel space. Since Long-LRM cannot process hundreds of generated frames, we uniformly subsample the GEN3C output before using it as input. Quantitative comparisons are reported in Table 4.

BTimer + GEN3C For the task of image to 3DGS generation, we show qualitative comparisons with BTimer (GEN3C) in Fig. 10 and qualitative results in Tab. 6a. We observe higher quality and less artifacts for our method.

Table 5: **CLIP and IQA metrics.** Semantic alignment evaluated with CLIP (ViT-B/32, ViT-B/16, ViT-L/14; higher is better) and perceptual image quality evaluated with NIQE (lower is better).

Method	CLIP-B/32 ↑	CLIP-B/16 ↑	CLIP-L/14 ↑	NIQE ↓
Long-LRM (GEN3C)	21.10	20.31	16.43	19.70
Ours	**32.58**	**32.38**	**27.24**	**11.34**

Table 6: **Quantitative results on static and dynamic *Lyra* datasets.**

Method	PSNR ↑	SSIM ↑	LPIPS ↓
BTimer (GEN3C)	16.32	0.580	0.427
Ours	**24.92**	**0.834**	**0.183**

(a) **Comparison on static *Lyra* dataset.**

Method	PSNR ↑	SSIM ↑	LPIPS ↓
BTimer (GEN3C)	20.29	0.687	0.315
Ours	**23.07**	**0.779**	**0.231**

(b) **Comparison on dynamic *Lyra* dataset.**

CLIP and IQA metrics. We evaluate our method using additional semantic and perceptual quality metrics. For semantic alignment, we report CLIP scores using three variants: ViT-B/32, ViT-B/16, and ViT-L/14, where higher values indicate better text–image alignment. For image quality assessment, we compute the Naturalness Image Quality Evaluator (NIQE), where lower values correspond to better perceptual quality. Tab. 5 reports the results and compares our method with the GEN3C + Long-LRM baseline. Our approach consistently achieves higher CLIP scores across all variants and substantially lower NIQE values, indicating improved semantic consistency and perceptual image quality.

D.3 Dynamic 3D Evaluation

BTimer (GEN3C) comparison. For dynamic 3D evaluation, we compare our method with BTimer (GEN3C), as discussed in Sec. D.1, on 100 out-of-distribution videos from our dynamic *Lyra* dataset. We crop the video to 512×512 for fair comparisons, as BTimer was mainly trained on that resolution. We show results in Tab. 6b and observe that our method significantly outperforms BTimer (GEN3C).

D.4 Depth Visualization

We visualize depths from our generated 3DGS in Fig. 11. Using depth supervision prevents flat geometries without sacrificing visual quality.

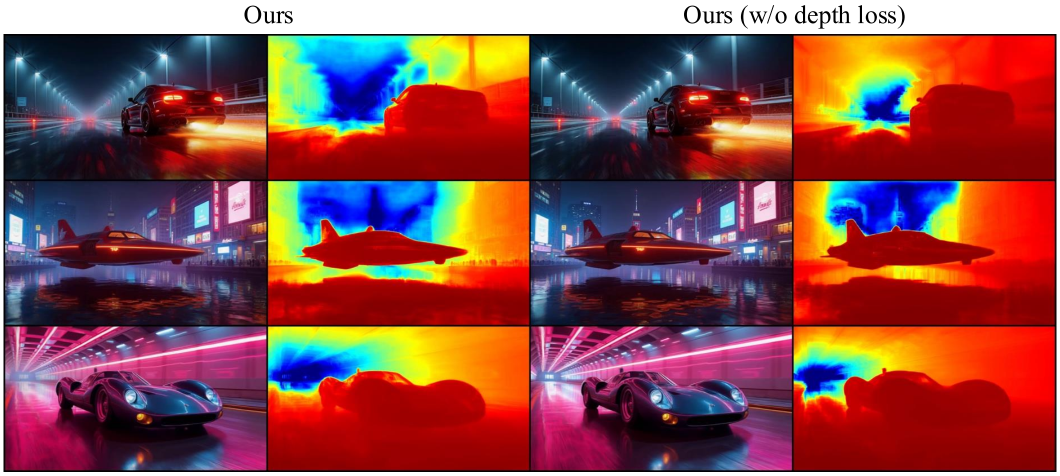
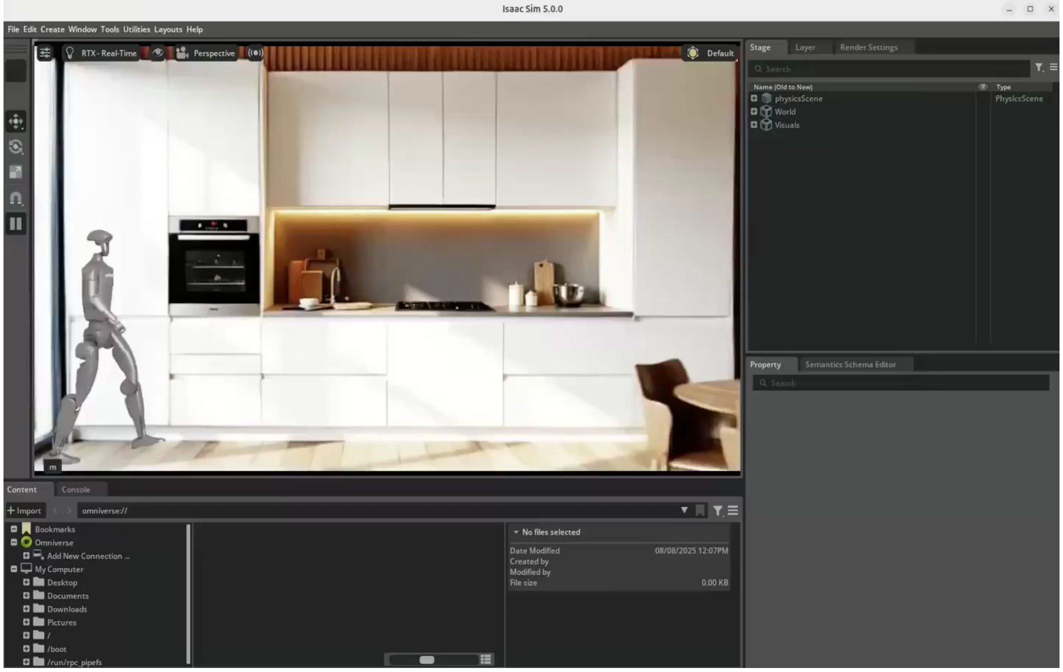


Figure 11: **Depth loss ablation.** We visualize 3DGS renderings and corresponding depth maps. Using the depth loss as additional supervision prevents flat geometries without sacrificing visual quality.



Date Modified 08/08/2025 12:07PM
Created by
Modified by
File size 0.00 KB

Figure 12: **Robot Simulation.** We visualize a frame of a robot simulation within the Isaac Sim 5.0 simulation framework that takes a generated 3DGS scene from our method as input.

E Robot Simulation

One of our primary motivations for building a diverse 3D generator is its ability to synthesize simulation environments for autonomous agents. To this end, we develop an end-to-end pipeline: we first generate 3D scenes from text with our model, export the 3D Gaussians as a .ply file, and convert the .ply files into a .usdz format. We use 3DGUT (Wu et al., 2025a) and its export function to create .usdz files. The resulting .usdz file can then be imported into the NVIDIA Isaac robot development platform as a physically based virtual environment. AI-based robots can be trained and tested under diverse conditions within these environments. We present early demo results on our supplementary webpage and a screenshot of the interface in Fig. 12.

F Related Work

In this section, we will discuss additional related work. We highlight the differences to previous works CAT3D (Gao et al., 2024b) and Bolt3D (Szymanowicz et al., 2025b) in Fig. 13.

3D generation. Early approaches to 3D generation primarily focused on single object categories, extending GANs into the 3D domain by leveraging neural renderers as an inductive bias (DeVries et al., 2021; Chan et al., 2022; Or-El et al., 2022; Schwarz et al., 2022; Bahmani et al., 2023b). With the introduction of CLIP-based supervision (Radford et al., 2021), the field advanced toward more flexible and diverse asset creation, enabling both text-driven generation and editing (Chen et al., 2018; Jain et al., 2022; Sanghi et al., 2022; Jetchev, 2021; Gao et al., 2023; Wang et al., 2022). More recently, diffusion models have significantly improved quality by replacing CLIP guidance with Score Distillation Sampling (SDS) (Poole et al., 2023; Wang et al., 2023b; Lin et al., 2023a; Chen et al., 2023b; Liang et al., 2023; Wang et al., 2023a; Li et al., 2024e; He et al., 2024b; Ye et al., 2024; Liu et al., 2024b; Yu et al., 2023; Katzir et al., 2024; Lee et al., 2024a; Sun et al., 2024a).

To further enhance structural coherence, several methods enforce multi-view consistency by generating scenes from different perspectives (Lin et al., 2023b; Liu et al., 2023; Shi et al., 2024; Feng et al., 2024a; Liu et al., 2024a; Kim et al., 2023; Voleti et al., 2024; Höllein et al., 2024; Tang et al., 2024c; Gao et al., 2024b; Wang et al., 2025f; Kant et al., 2025; Yuan et al., 2025; Ren et al., 2024b), while others adopt iterative inpainting as a strategy for scene synthesis (Höllein et al., 2023; Shriram et al., 2024). Another line of work explores lifting 2D images into 3D using NeRF (Mildenhall et al., 2020),

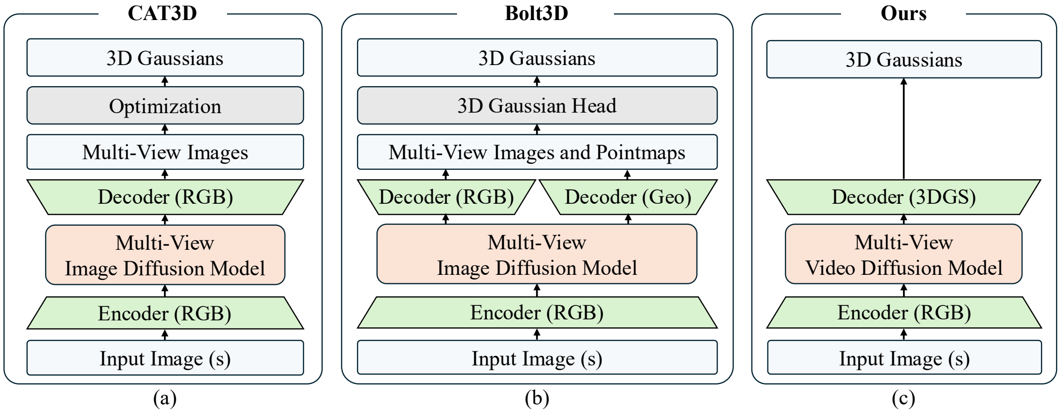


Figure 13: **Approaches for 3D Generation.** CAT3D (Gao et al., 2024b) proposed a multi-view image diffusion model that outputs images from novel viewpoints; subsequently, the images are reconstructed into 3D Gaussians using optimization. Bolt3D (Szymanowicz et al., 2025b) fine-tunes CAT3D to output pointmaps using a geometry autoencoder; instead of relying on optimization, the 3D scene is predicted with a feed-forward 3D Gaussian head. In contrast, our work builds upon a pre-trained camera-controlled video diffusion model and *directly* decodes the multi-view video latents into 3D Gaussians.

3D Gaussian Splatting (Kerbl et al., 2023), or mesh-based representations coupled with diffusion models (Chan et al., 2023; Tang et al., 2023a; Gu et al., 2023; Liu et al., 2024c; Yoo et al., 2023; Tewari et al., 2023; Qian et al., 2024b; Long et al., 2024; Wan et al., 2024; Szymanowicz et al., 2023; Lu et al., 2024).

Feed-forward 3D models. Complementary efforts investigate fast, feed-forward techniques that directly predict 3D content from images or text inputs (Hong et al., 2024; Li et al., 2024b; Xu et al., 2024e;d; Zhang et al., 2024a; Han et al., 2024; Jiang & Wang, 2024; Xie et al., 2024a; Tang et al., 2024b; Tochilkin et al., 2024; Qian et al., 2024a; Szymanowicz et al., 2024; 2025a; Liang et al., 2025a; Szymanowicz et al., 2025b; Schwarz et al., 2025; Yang et al., 2025; Zhang et al., 2025a). However, in contrast to our method, these approaches remain limited to producing static 3D scenes. Other methods focus on specific scenes such as faces (Kirschstein et al., 2025). Concurrent work (Liang et al., 2025b; Xu et al., 2025) tackles real scenes but can not handle diverse generated scenes or large viewpoint changes.

4D generation. Recent years have seen rapid progress in 4D generation, i.e., dynamic 3D scene synthesis. Many approaches leverage input text prompts or images as guidance. Following the introduction of large-scale generative models for this task (Singer et al., 2023), subsequent works have made notable advances in both visual fidelity and motion quality (Ren et al., 2023; Ling et al., 2024a; Bahmani et al., 2024b; Zheng et al., 2024b; Gao et al., 2024a; Yang et al., 2024a; Jiang et al., 2024b; Miao et al., 2024; Li et al., 2024a; Zhang et al., 2024g; Yuan et al., 2024; Jiang et al., 2024a).

While text conditioning is widely used, alternative methods aim to lift 2D images or videos into dynamic 3D scenes (Ren et al., 2023; Zhao et al., 2023; Yin et al., 2023; Pan et al., 2024; Zheng et al., 2024b; Ling et al., 2024a; Gao et al., 2024a; Zeng et al., 2024b; Chu et al., 2024; Wu et al., 2024; Yang et al., 2024b; Wang et al., 2024c; Feng et al., 2024b; Sun et al., 2024c; Zhang et al., 2024c; Ren et al., 2024a; Lee et al., 2024b; Li et al., 2024c; Van Hoorick et al., 2024; Uzolas et al., 2024; Chai et al., 2024; Liang et al., 2024; Li et al., 2024d; Xie et al., 2024b; Li et al., 2025a; Wang et al., 2025e; Zhu et al., 2025; Zhou et al., 2025; Hu et al., 2025; Park et al., 2025). Other efforts explore incorporating physics priors into generation pipelines (Lin et al., 2024; Huang et al., 2024b; Zhang et al., 2024f; Kiray et al., 2025), or enhancing motion controllability with template-driven approaches (Zhang et al., 2024d; Sun et al., 2024b; Chen et al., 2024b). A parallel direction emphasizes compositional and interactive 4D generation (Bahmani et al., 2024a; Xu et al., 2024b; Cao et al., 2024; Yu et al., 2024a; Zeng et al., 2024a; Zhu et al., 2024).

Another strand of research extends 3D GAN frameworks into the 4D setting by training directly on 2D video data (Bahmani et al., 2023a; Xu et al., 2023). However, such methods are typically limited by small-scale, single-category datasets, which constrain generalization. Moreover, most existing approaches remain object-centric, often neglecting background elements. As a result, their overall visual fidelity lags behind the high photorealism achieved by recent video generation models, which our method leverages.

Camera-conditioned video models. Recent progress has focused on incorporating camera control into video diffusion models. The pioneering work MotionCtrl (Wang et al., 2024e) introduced camera conditioning by augmenting pre-trained video models (Chen et al., 2023a; Blattmann et al., 2023) with extrinsic matrices. Subsequent efforts (He et al., 2024a; Xu et al., 2024c; Kuang et al., 2024; Ju et al., 2025; He et al., 2025; Li et al., 2025c; Wang et al., 2025c; 2024d; Lei et al., 2025; Liu et al., 2025) enhanced conditioning by representing cameras using Plücker coordinates, while others use follow MotionCtrl to simply use poses flattened (Wang et al., 2025b). Another line of research (Hu et al., 2024; Xiao et al., 2024; Ling et al., 2024c; Hou et al., 2024) enabled camera motion control without introducing additional trainable parameters. Despite these advances, all of the above rely on U-Net-based architectures. Other approaches use low-rank approaches for camera control (Zhang et al., 2025b).

Several recent works continue to extend U-Net-based approaches for camera control (Zhao et al., 2024; Xu et al., 2024a; Zheng et al., 2024a; Zhang et al., 2024b; Yu et al., 2024b; Lei et al., 2024), while others begin exploring diffusion transformers. Cheong et al.(Cheong et al., 2024) propose one such transformer-based framework, though its scene and visual quality remain limited. DimensionX(Sun et al., 2024d) introduces joint space-time control in diffusion transformers, but relies on pre-defined, non-continuous camera trajectories. Other works (Huang et al., 2024a; Mai et al., 2025; Jiang et al., 2025b) instead explore pose and geometry estimation with a video DiT. CAT4D (Wu et al., 2025b) proposes a multi-view video diffusion model derived from fine-tuning a multi-view backbone. Large-scale synthetic data generation further boosts performance, as demonstrated by SynCamMaster (Bai et al., 2025b) and ReCamMaster (Bai et al., 2025a), which achieve strong results on camera-controlled video generation using transformer-based architectures. Finally, 4DiM (Watson et al., 2024) trains a space-time diffusion model from scratch for novel view synthesis from a single input image. Another line of work edits camera movements of videos with underlying motion using camera retargeting (Seo et al., 2025; Ma et al., 2025). One popular direction has been to use a spatio-temporal cache, e.g., using point clouds or optical flow, to condition the camera-controlled video generation (Wu et al., 2025c; Ren et al., 2025; Wang et al., 2025d; Cao et al., 2025; YU et al., 2025; Jin et al., 2025; Xing et al., 2025; Yan et al., 2025; Li et al., 2025b; Gu et al., 2025).