

# DIFF4SPLAT: CONTROLLABLE 4D SCENE GENERATION WITH LATENT DYNAMIC RECONSTRUCTION MODELS

Panwang Pan<sup>\*†‡</sup>, Chenguo Lin<sup>\*</sup>, Jingjing Zhao, Chenxin Li, Yuchen Lin, Haopeng Li, Honglei Yan, Kairun Wen, Yunlong Lin, Yixuan Yuan, Yadong Mu<sup>‡</sup>

Peking University, The Chinese University of Hong Kong, Xiamen University

<sup>\*</sup> Equal contribution <sup>†</sup> Project lead <sup>‡</sup> Corresponding author

<https://paulpanwang.github.io/Diff4Splat>

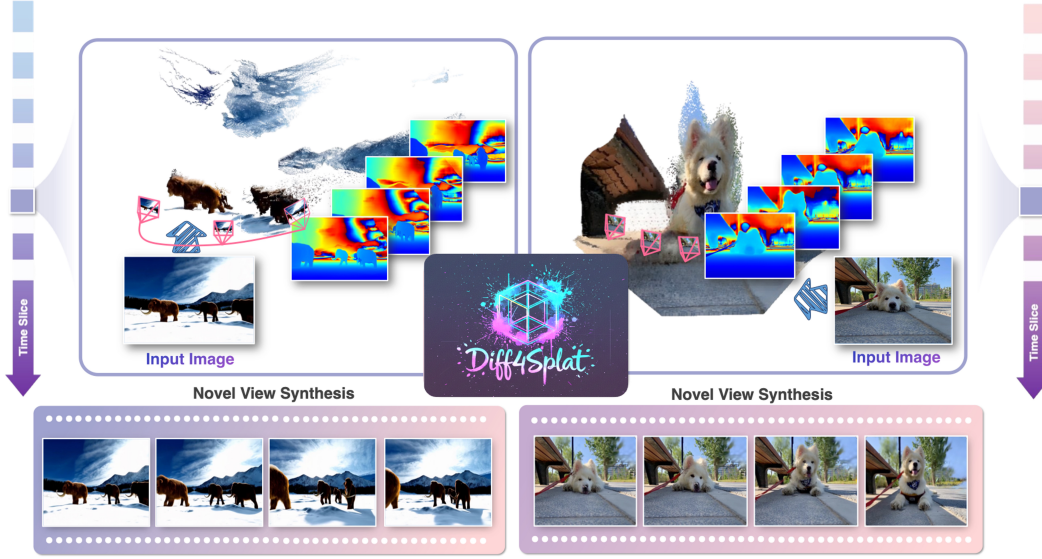


Figure 1: Given a single image, a specified camera trajectory, and an optional text prompt, our diffusion-based framework directly generates a **deformable 3D Gaussian field without test-time optimization**. The resulting representation supports diverse applications, including video generation, depth rendering, and novel view synthesis, enabling real-time rendering of dynamic scenes and interactive virtual exploration.

## ABSTRACT

We introduce **DIFF4SPLAT**, a **feed-forward method** that synthesizes controllable and explicit 4D scenes **from a single image**. Our approach unifies the generative priors of video diffusion models with geometry and motion constraints learned from large-scale 4D datasets. Given a single input image, a camera trajectory, and an optional text prompt, **DIFF4SPLAT** directly predicts a deformable 3D Gaussian field that encodes appearance, geometry, and motion, all in a single forward pass, without test-time optimization or post-hoc refinement. At the core of our framework lies a video latent transformer, which augments video diffusion models to jointly capture spatio-temporal dependencies and predict time-varying 3D Gaussian primitives. Training is guided by objectives on appearance fidelity, geometric accuracy, and motion consistency, enabling **DIFF4SPLAT** to synthesize high-quality 4D scenes in 30 seconds. We demonstrate the effectiveness of **DIFF4SPLAT** across video generation, novel view synthesis, and geometry extraction, where it matches or surpasses optimization-based methods for dynamic scene synthesis while being significantly more efficient.

## 1 INTRODUCTION

Recent advances in monocular 4D reconstruction have shown promising results, yet their practicality is often limited by lengthy optimization (Wu et al., 2024b; Lei et al., 2024) or a lack of flexibility (Liang et al., 2024c; Shen et al., 2025). Existing approaches to controllable 4D scene generation from a single image typically decompose the task into progressive video generation (Ren et al., 2025) followed by 3D neural reconstruction (Kerbl et al., 2023b; Yang et al., 2024b). While effective, these pipelines rely on multiple non-differentiable modules, require costly test-time optimization, or restrict themselves to static 3D scenes (Liang et al., 2024a) due to dataset limitations. More recently, the reliance on dynamic pointmaps in recent feed-forward methods (Zhu et al., 2025; Chen et al., 2025b) limits rendering quality, precluding photorealism and frequently introducing holes and artifacts. These challenges call for a unified and efficient framework that can directly generate dynamic 3DGS content.

We tackle the challenging task of single-stage controllable 4D scene generation from a single image, which requires simultaneous camera pose control, metric-scale geometry and motion prediction, and photorealistic rendering, all within a holistic deformable 3D particle-based representation (Yang et al., 2023). This problem is inherently ill-posed: it demands not only realistic image synthesis but also the recovery of dynamic geometry from sparse conditioning signals. The difficulty is further compounded by the scarcity of real-world video datasets with metric-scale depth. A successful solution must therefore combine photorealistic rendering with spatio-temporal coherence in the generated content. Such capabilities would unlock a wide range of applications, including immersive XR content creation, realistic environments for robotics, and scalable autonomous driving simulation.

As illustrated in Fig. 1, we aim to build a unified framework that directly predicts full 4D representations without test-time optimization or post-processing steps. This enables video generation, depth rendering, and novel view synthesis within a single diffusion model. To this end, we introduce **DIFF4SPLAT**, a holistic 4D diffusion transformer designed for scalable, data-driven scene generation. Addressing the scarcity of physically-grounded 4D data, we construct a large-scale annotation pipeline that converts real-world videos into spatio-temporal pointmaps with metric depth. To capture both visual and spatio-temporal dependencies, we extend the diffusion backbone with a **Latent Dynamic Reconstruction Model**, which transforms latent 2D tokens under temporal and camera embeddings. A lightweight prediction head then decodes these tokens into deformable 3D Gaussians, enabling real-time rendering of novel views and geometric maps such as depth. Our dataset provides rich supervision over appearance, geometry, and motion. Leveraging it, **DIFF4SPLAT** achieves state-of-the-art efficiency and geometric fidelity, while its unified representation yields significantly improved motion quality.

Our contributions can be summarized as follows:

- We propose **DIFF4SPLAT**, a unified diffusion-based model that directly generates deformable 3D Gaussians for controllable 4D scene synthesis.
- We construct a large-scale 4D dataset from synthetic and in-the-wild videos, annotated with appearance, metric-scale geometry, and motion.
- Extensive experiments demonstrate that **DIFF4SPLAT** produces high-fidelity 4D scenes from a single image, outperforming two-stage pipelines and existing camera-controlled video generation methods in both quality and efficiency.

## 2 RELATED WORK

**Video Diffusion Models** Video diffusion models (Ho et al., 2022) have demonstrated a remarkable capacity for generating high-quality, temporally coherent videos. Fine-grained control is typically achieved by adapting conditional image synthesis strategies (Zhang et al., 2023; Mou et al., 2024; Li et al., 2023b) to the video domain, incorporating diverse signals such as RGB images (Blattmann et al., 2023a; Xing et al., 2023; 2024a), depth maps (Xing et al., 2024b; Esser et al., 2023), motion trajectories (Yin et al., 2023; Niu et al., 2024), and semantic maps (Peruzzo et al., 2024). Despite these advancements, explicit camera motion control remains a relatively underexplored area. Existing approaches often rely on predefined motion categories (Guo et al., 2023; Blattmann et al., 2023a) or learnable LoRA modules (Hu et al., 2022). While methods like MotionCtrl (Wang et al., 2024b) employ camera extrinsics, they exhibit limited precision in complex scenarios, and MultiDiff (Müller

et al., 2024) is constrained by class-specific training. More recently, several works (Xu et al., 2024; He et al., 2024; 2025) have leveraged Plücker coordinates (Sitzmann et al., 2021) for camera control, but still face challenges in producing realistic video outputs. Notably, the majority of current research generates videos as 2D frame sequences, largely overlooking the joint generation of dynamic 3D representations (e.g., dynamic 3DGS).

**Static 3D Scene Generation** Recent progress in generative models (Ho et al., 2020; Rombach et al., 2022b; Yang et al., 2025; Wang et al., 2025a) and 3D representations (Kerbl et al., 2023a; Mildenhall et al., 2020) has significantly advanced static 3D scene generation. One prominent research direction focuses on structured scene generation from layouts or graphs (Gao et al., 2024; Bai et al., 2023a; Po & Wetzstein, 2024; Vilesov et al., 2023; Yuan et al., 2025; Lin et al., 2025b; Lin & Mu, 2024; Lin et al., 2024a). Another line of research, more related to our work, addresses open-world scene generation from weak conditioning signals like text (Chung et al., 2023; Zhou et al., 2024) or images (Chung et al., 2023; Yu et al., 2024b; Liang et al., 2024a). These methods often rely on image diffusion models (Ho et al., 2020; Rombach et al., 2022b) as a backbone to provide strong 3D priors (Chung et al., 2023; Zhou et al., 2024; Yu et al., 2024b; Szymanowicz et al., 2025; Lin et al., 2025a; Wewer et al., 2024). The rise of video diffusion models has also motivated studies (Liang et al., 2024a; Liu et al., 2024; Yu et al., 2024c; Sun et al., 2024a) to leverage them for improved 3D-aware consistency. Our work distinguishes itself by pioneering dynamic scene generation, addressing the critical challenge of modeling motion.

**Dynamic 4D Scene Generation** Static 3D generation methods are inherently limited to motionless scenes. The natural, albeit challenging, progression is dynamic 4D scene generation (Zhao et al., 2024b; Zhang et al., 2024; Chu et al., 2024; Liang et al., 2024d; Lin et al., 2025c; Li et al., 2024; Wang et al., 2025c; Zhu et al., 2025). Due to dataset limitations (Zhou et al., 2018a; Dai et al., 2017; Yeshwanth et al., 2023; Ling et al., 2024; Yu et al., 2023), prior works often tackle sub-problems. Some methods require a video and multi-view images of the first frame (Yu et al., 2024a; Wang et al., 2024a; Xie et al., 2024). Others generate 4D Gaussian Splatting from monocular video (Chu et al., 2025; Wu et al., 2024b; Liang et al., 2024d; Shen et al., 2025) or rely on costly per-scene optimization (Lei et al., 2024; Li et al., 2023c; Zhao et al., 2024a; Wang et al., 2025b; Wu et al., 2024a; Sun et al., 2024c). Recent feed-forward works generate dynamic pointmaps (Zhu et al., 2025; Chen et al., 2025b), but this kind of representation struggles to achieve photorealism, resulting in renderings with holes and artifacts. In contrast, our work introduces a generalizable method that generates an **explicit deformation Gaussian field** from a single image, without per-scene optimization.

### 3 METHODOLOGY

Our primary objective is the generation of a dynamic 4D scene representation from a single input image  $\mathbf{I}_0 \in \mathbb{R}^{H \times W \times 3}$ , text prompt  $\mathbf{C}_{\text{ctx}}$ , and the corresponding camera poses represented by Plücker embeddings (Jia, 2020)  $\mathcal{P} \in \mathbb{R}^{T \times H \times W \times 6}$ . As shown in Fig. 2, our methodology integrates a video diffusion model with a novel latent reconstruction Transformer. This unified framework synergistically combines 2D appearance priors, geometric constraints, and motion cues to synthesize high-fidelity 4D scenes. First, we leverage a pre-trained video diffusion model, conditioned on camera poses and the input image, to produce a video latent tensor  $\mathbf{z} \in \mathbb{R}^{n \times h \times w \times c}$ , where  $n$  is the number of synthesized latent features, and  $h, w, c$  denote the height, width, and channel dimensions of the latent features, respectively. We then introduce a Latent Dynamic Reconstruction Model (Sec. 3.2) that effectively integrates camera conditions with the generated latent features to predict a deformable Gaussian field, enabling rendering at novel viewpoints and time instances. Second, to facilitate dynamic scene generation, we augment the foundational static 3D Gaussian Splatting representation (Kerbl et al., 2023a) with an efficient mechanism for inter-frame deformation (Sec. 3.3). Third, we introduce a unified supervision scheme (Sec. 3.4) that incorporates photometric, geometric, and motion losses. Finally, we devise a progressive training strategy to ensure high-fidelity texture synthesis and enforce robust geometric constraints.

#### 3.1 DATA CURATION

We start by developing a scalable 4D data annotation pipeline, meticulously designed to convert real-world videos into spatio-temporal point maps at metric scales. Our data curation strategy systematically integrates two complementary types of data sources:

① *Synthetic Datasets*: We leverage seven synthetic datasets: TartanAir (Wang et al., 2020), Matrix-City (Li et al., 2023a), PointOdyssey (Zheng et al., 2023), DynamicReplica (Karaev et al., 2023), Spring (Mehl et al., 2023), VKITTI2 (Cabon et al., 2020), and MultiCamVideo (Bai et al., 2025). These datasets provide precise ground-truth annotations and controlled environmental variations, which are essential for learning robust geometric priors. ② *Real-world Datasets*: We incorporate two real-world datasets: RealEstate10K (Zhou et al., 2018b) and Stereo4D (Jin et al., 2025). These datasets offer authentic scene complexity and natural variations, which are crucial for enhancing the model’s generalization capabilities. Inspired by (Zhu et al., 2025), we employ VideoDepthAnything (Chen et al., 2025a) and MegaSaM (Li et al., 2025b) to recover metric scale from these datasets, enabling more precise camera control within our generative framework (Bahmani et al., 2024).

Through this comprehensive data collection and processing pipeline, we amass approximately 130,000 high-quality 4D training scenes. Following a rigorous quality control protocol, which includes dynamic object masking and reprojection error filtering. We curate a refined dataset of approximately 100,000 synchronized multi-view videos, each annotated with metric point-maps and point motion trajectories. More technical details are available in Appendix B.

### 3.2 LATENT DYNAMIC RECONSTRUCTION MODEL

While video diffusion models have demonstrated remarkable success in generating high-quality visual content, their direct application to synthesizing 3D-aware latents is non-trivial. This challenge arises from their inherent lack of explicit control over camera pose trajectories and their propensity to generate dynamic content that may lack the consistency required for robust 3D reconstruction. Drawing inspiration from recent advancements in latent-based diffusion models (Blattmann et al., 2023b; Rombach et al., 2022a; Pan et al., 2024; Liang et al., 2024b), we introduce the **Latent Dynamic Reconstruction Model (LDRM)**, which significantly mitigates the computational overhead associated with per-scene optimization strategies. LDRM utilizes a pre-trained video diffusion model, conditioned on camera poses and an input image, to generate the latent tensor  $\mathbf{z}$ . The resulting video latents are inherently compact and 3D-aware, encapsulating a multi-view representation of the scene that is consistent in both structure and appearance, rendering them ideal for subsequent 3D lifting. Given the video latent tensor  $\mathbf{z} \in \mathbb{R}^{n \times h \times w \times c}$  and the corresponding camera poses, we first transform these inputs into latent and pose tokens. Patchify modules ensure that both token sets possess identical sequence lengths. These token sets are then concatenated channel-wise and subsequently processed by a series of Transformer blocks (Ainslie et al., 2023). A lightweight decoding module regresses the attributes of 3D Gaussians from the Transformer’s output tokens and uses a 3D deconvolutional layer to establish a pixel-level correspondence with the source video frames.

### 3.3 DEFORMABLE GAUSSIAN FIELDS

A static 3D scene can be represented as a collection of  $M$  Gaussian primitives  $\{\mathbf{G}_p\}_{p=1}^M$ . Each Gaussian  $\mathbf{G}_p$  is characterized by its mean location  $\boldsymbol{\mu}_p \in \mathbb{R}^3$ , scaling factors  $\mathbf{s}_p \in \mathbb{R}^3$ , orientation quaternion  $\mathbf{q}_p \in \mathbb{R}^4$ , opacity  $\alpha_p \in \mathbb{R}$ , and color features  $\mathbf{c}_p \in \mathbb{R}^C$ . We use Spherical Harmonics (SH) to model view-dependent effects. The spatial influence of each Gaussian is given by:

$$\mathbf{G}_p(\mathbf{x}) := \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_p)^\top \boldsymbol{\Sigma}_p^{-1}(\mathbf{x} - \boldsymbol{\mu}_p)\right), \quad (1)$$

where  $\boldsymbol{\Sigma}_p$  is the covariance matrix derived from  $\mathbf{s}_p$  and  $\mathbf{q}_p$ . Inspired by (Yang et al., 2024b; Lin et al., 2024b; Liang et al., 2025), we introduce a deformable 3D Gaussian formulation to represent dynamic scene. For each Gaussian  $p$  at time step  $t$ , the predicted deformation field comprises a displacement for its mean,  $\Delta\boldsymbol{\mu}_p^t \in \mathbb{R}^3$ ; an adjustment to its rotation,  $\Delta\mathbf{q}_p^t \in \mathbb{R}^4$ ; and a modification to its scale,  $\Delta\mathbf{s}_p^t \in \mathbb{R}^3$ . The deformed parameters at time  $t$  are updated as follows:  $\boldsymbol{\mu}_p^t := \boldsymbol{\mu}_p^0 + \Delta\boldsymbol{\mu}_p^t$ ,  $\mathbf{q}_p^t := \mathbf{q}_p^0 \otimes \Delta\mathbf{q}_p^t$  (quaternion multiplication), and  $\mathbf{s}_p^t := \mathbf{s}_p^0 + \Delta\mathbf{s}_p^t$ . These deformed Gaussians are then rendered using a differentiable Gaussian rasterization pipeline. Deformable Gaussian Fields is equipped with the LDRM, which generates a Gaussian feature map  $\mathbf{G} \in \mathbb{R}^{(T \times H \times W) \times K_g}$ , where  $K_g$  denotes the number of parameters for each Gaussian primitive. Concurrently, the LDRM predicts a corresponding deformation map  $\mathbf{D} \in \mathbb{R}^{(T \times H \times W) \times K_d}$ . The dimensionality of this deformation,  $K_d = 10$ , comprises offsets for the mean ( $\Delta\boldsymbol{\mu} \in \mathbb{R}^3$ ), rotation ( $\Delta\mathbf{q} \in \mathbb{R}^4$ ), and scale ( $\Delta\mathbf{s} \in \mathbb{R}^3$ ).



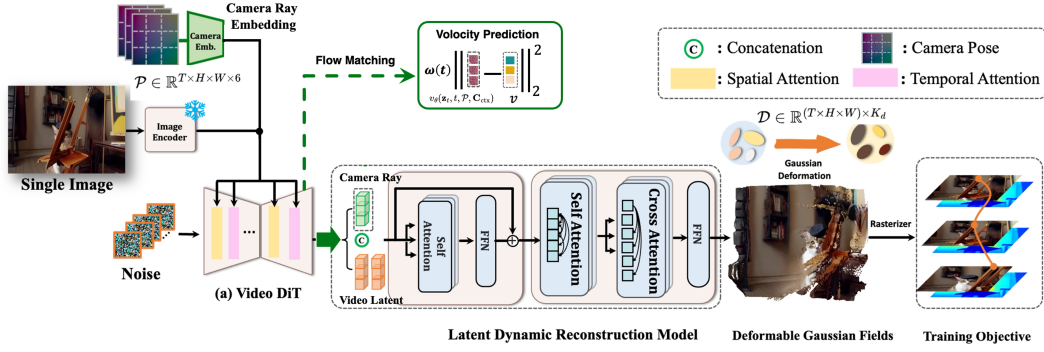


Figure 2: **Architecture of DIFF4SPLAT.** We present a high-fidelity dynamic 3DGS generation method from a single image through four key innovations: (1) video diffusion latents processed by our novel Transformer (Sec. 3.2), (2) a dynamic 3DGS deformation mechanism (Sec. 3.3), (3) unified supervision with photometric, geometric, and motion losses (Sec. 3.4), and (4) a progressive training scheme for robust geometry and texture.

### 3.4 TRAINING OBJECTIVE

To enhance the geometric consistency of the generated latents, we introduce a progressive training scheme that jointly optimizes the network across multi-tasks via differentiable rendering.

**Flow Matching Loss** The Flow Matching (FM) (Lipman et al., 2023) approach learns the vector field that transports a noise distribution to the data distribution. Let  $\mathbf{z}^{(0)}$  be a clean latent sequence from the data distribution  $p_{\text{data}}$ , and  $\mathbf{z}^{(1)} \sim \mathcal{N}(0, \mathbf{I})$  be a sample from the prior gaussian noise. A probability path  $\mathbf{p}_t(\mathbf{z}|\mathbf{z}^{(0)}, \mathbf{z}^{(1)})$  connects these samples, typically via linear interpolation  $\mathbf{z}_t = (1-t)\mathbf{z}^{(0)} + t\mathbf{z}^{(1)}$  for  $t \in [0, 1]$ . The corresponding target vector field is  $u_t(\mathbf{z}_t|\mathbf{z}^{(0)}, \mathbf{z}^{(1)}) = \mathbf{z}^{(1)} - \mathbf{z}^{(0)}$ . Our model,  $v_\theta(\mathbf{z}_t, t, \mathcal{P}, \mathbf{C}_{\text{ctx}})$ , is trained to approximate this vector field by minimizing:

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t, \mathbf{z}^{(0)}, \mathbf{z}^{(1)}, \mathcal{P}, \mathbf{C}_{\text{ctx}}} \left[ w(t) \| v_\theta(\mathbf{z}_t, t, \mathcal{P}, \mathbf{C}_{\text{ctx}}) - (\mathbf{z}^{(1)} - \mathbf{z}^{(0)}) \|_2^2 \right], \quad (2)$$

where  $w(t)$  is a weighting function for different noise levels and conditioning information (text prompt  $\mathbf{C}_{\text{ctx}}$  and Plücker embeddings  $\mathcal{P}$ ) is incorporated into  $v_\theta$ .

**Photometric Loss** To facilitate high-quality novel view synthesis, we optimize the 3DGS parameters using a composite loss:

$$\mathcal{L}_{\text{photo}} = \text{MSE}(\hat{\mathbf{I}}^k, \mathbf{I}^k) + \lambda_p \cdot \text{LPIPS}(\hat{\mathbf{I}}^k, \mathbf{I}^k), \quad (3)$$

where  $\hat{\mathbf{I}}^k$  is the rendered image for view  $k$ ,  $\mathbf{I}^k$  is the ground-truth image, and  $\lambda_p$  is a balancing coefficient for the LPIPS (Zhang et al., 2018) term.

**Geometric Loss** Inspired by (Li et al., 2025a), we introduce a geometric regularization term to enforce accurate depth relationships. Let  $\hat{D}_k$  be the rendered depth map and  $D_k^*$  be the ground-truth depth for view  $k$ .

$$\mathcal{L}_{\text{geo}}(\hat{D}_k, D_k^*) = 1 - \frac{\text{Cov}(\hat{D}_k, D_k^*)}{\sqrt{\text{Var}(\hat{D}_k)\text{Var}(D_k^*)}}, \quad (4)$$

where  $\text{Cov}$  and  $\text{Var}$  are covariance and variance functions. We also apply a total variation loss,  $\mathcal{L}_{\text{TV}} = \|\nabla \hat{D}_k\|_1$ , to enforce local smoothness.

**Motion Loss** Given 3D point tracking data, the ground-truth motion for a point  $j$  is its displacement  $\Delta \mathbf{x}_j$ . The motion loss is:

$$\mathcal{L}_{\text{motion}} = \frac{1}{|\mathcal{O}|} \sum_{j \in \mathcal{O}} (\lambda_m \|\Delta \hat{\mathbf{x}}_j - \Delta \mathbf{x}_j\|_2 + \|\Delta \hat{\mathbf{x}}_j\|_1), \quad (5)$$

where  $\mathcal{O}$  is the set of tracked points,  $\Delta \hat{\mathbf{x}}_j$  is the predicted displacement, and  $\lambda_m$  is a weighting coefficient.

**Progressive Training Scheme** To bridge the domain gap between video latents and the 3DGS representation, we introduce a three-stage progressive training scheme.

❶ **Static Geometry Pre-training (40K iterations).** We first establish a strong geometric prior by training LDRM on static scenes (e.g., TartanAir, RealEstate10K) at a low resolution ( $256 \times 256$ ), using only photometric and geometric losses. During this stage, the deformation module (an 8-layer DPT head) is frozen.

❷ **High-Resolution Refinement (40K iterations).** With the deformation module still frozen, we enhance reconstruction fidelity by training on static scenes under a high resolution ( $512 \times 512$ ).

❸ **Dynamic Scene Fine-tuning (20K iterations).** Finally, we unfreeze and fine-tune *the entire model* on dynamic datasets (PointOdyssey, DynamicReplica, Spring, VKITTI2, and Stereo4D). This stage employs the complete loss function, including a motion loss term, to learn temporal deformations. This progressive strategy, combined with our large-scale 4D dataset, enables our model to learn complex dynamics and generate high-fidelity, temporally coherent 4D scenes.

## 4 EXPERIMENTAL EVALUATION

### 4.1 IMPLEMENTATION DETAILS

Our framework builds upon a pretrained Video Diffusion Transformer model, CogVideoX (Yang et al., 2024a), operating within the latent space of a 32-channel  $4 \times 8 \times 8$  compression 3D Causal Variational Autoencoder. The architecture comprises 32 blocks with a hidden dimensionality of 4096, specifically designed for image-to-deformation Gaussian field generation. Our LDRM architecture is composed of 16 standard Transformer blocks, the latent features have a channel dimension of  $c = 32$  and are projected into a 64-dimensional embedding space before being processed by the Transformer backbone. To enable text control capabilities, each DiT block incorporates a cross-attention layer that integrates image embedding information from the T5 model (Raffel et al., 2020). For training, we employ the AdamW optimizer (Loshchilov & Hutter, 2019) with an initial learning rate of  $10^{-5}$  and a weight decay of  $10^{-4}$ . The loss weighting hyperparameters are set to  $\lambda_p = 0.5$  for the photometric loss and  $\lambda_m = 2$  for the motion loss. A cosine learning rate scheduler is utilized, and the model is trained for 100,000 iterations until convergence. This training process requires approximately 7 days on a setup of 32 A100 GPUs, using BF16 mixed precision. At inference time, our Deformable Gaussian Diffusion model generates a complete 4D scene in **30 seconds**.

### 4.2 EVALUATION PROTOCOL

**Baselines** We compare our holistic pipeline against the two-stage pipeline, which incorporates state-of-the-art techniques. Specifically, for this two-stage approach, we use AC3D (Bahmani et al., 2024) for single-image controllable video generation and Mosca (Lei et al., 2024) for dynamic Gaussian reconstruction. For comprehensive evaluation of camera controllability, we generated **160 evaluation samples** by applying five distinct camera trajectories (spiral, forward, backward, upward, and downward) to 32 unique text-captioned scenes.

**Metrics** Our evaluation encompasses both prompt-scene consistency and aesthetic quality through: CLIP similarity score (Radford et al., 2021), Aesthetic score (CLIP-Aesthetic) (Schuhmann, 2023), VLM-based visual scorer Q-Align (QA-Quality) (Wu et al., 2023), and video quality metrics: FVD (Unterthiner et al., 2019) and KVD (Unterthiner et al., 2018). For geometric integrity assessment, we employ the MAST3R (Leroy et al., 2024) algorithm for local correspondence matching between input views and generated novel views and provide metrics through: Average matching correspondences, subject consistency score, and background consistency score (Zheng et al., 2025). More details are provided in Appendix D.

### 4.3 QUANTITATIVE AND QUALITATIVE EVALUATION

**Quantitative Results** As shown in Tab. 1 and Tab. 2, our approach achieves competitive or superior performance across a variety of evaluation metrics. In terms of video generation and aesthetic quality (Tab. 1), our method delivers highly competitive results. Moreover, it significantly reduces reconstruction time to approximately 30 seconds. It offers a substantial efficiency improvement over methods like “AC3D + Mosca”, which require around **45 minutes**, while maintaining strong

Table 1: Quantitative comparison on appearance fidelity and aesthetic quality. † indicates that this method requires per-scene optimization. Best results are in **bold**.

| Method                                                   | Video Generation & Aesthetic Quality |              |               |                  |              | Rec. Time ↓ |
|----------------------------------------------------------|--------------------------------------|--------------|---------------|------------------|--------------|-------------|
|                                                          | FVD ↓                                | KVD ↓        | CLIP-Score ↑  | CLIP-Aesthetic ↑ | QA-Quality ↑ |             |
| <i>Camera-Controlled Video Generation</i>                |                                      |              |               |                  |              |             |
| CameraCtrl (He et al., 2024)                             | 478.192                              | 8.105        | 19.365        | 2.965            | 1.894        | 20s         |
| AC3D (Bahmani et al., 2024)                              | 339.431                              | 6.342        | 20.673        | 3.324            | 2.158        | 28s         |
| <i>Explicit 3DGS Representation</i>                      |                                      |              |               |                  |              |             |
| AC3D + Shape of Motion <sup>†</sup> (Wang et al., 2025b) | 373.045                              | 6.511        | 16.201        | 3.043            | 1.838        | 18min       |
| AC3D + SaV <sup>†</sup> (Sun et al., 2024b)              | 327.122                              | 5.816        | 19.018        | 4.371            | 2.382        | 35min       |
| AC3D + Mosca <sup>†</sup> (Lei et al., 2024)             | 235.961                              | <b>2.012</b> | 20.214        | 4.999            | <b>2.842</b> | 45min       |
| <b>Ours</b>                                              | <b>210.153</b>                       | <u>2.316</u> | <b>23.123</b> | <b>5.231</b>     | <u>2.813</u> | <b>30s</b>  |

Table 2: Quantitative comparison on geometric integrity and reconstruction time. † indicates that this method requires per-scene optimization. Best results are in **bold**.

| Method                                                   | Geometric Integrity |                             |                                | Rec. Time ↓ |
|----------------------------------------------------------|---------------------|-----------------------------|--------------------------------|-------------|
|                                                          | Avg. Matches ↑      | Subject Consistency Score ↑ | Background Consistency Score ↑ |             |
| <i>Camera-Controlled Video Generation</i>                |                     |                             |                                |             |
| CameraCtrl (He et al., 2024)                             | 2015.82             | 72.25                       | 74.53                          | 20s         |
| AC3D (Bahmani et al., 2024)                              | 2489.16             | 75.64                       | 75.91                          | 28s         |
| <i>Explicit 3DGS Representation</i>                      |                     |                             |                                |             |
| AC3D + Shape of Motion <sup>†</sup> (Wang et al., 2025b) | 2874.22             | 83.13                       | 83.33                          | 18min       |
| AC3D + SaV <sup>†</sup> (Sun et al., 2024b)              | 3035.43             | 85.96                       | 84.23                          | 35min       |
| AC3D + Mosca <sup>†</sup> (Lei et al., 2024)             | <u>4500.68</u>      | <b>86.23</b>                | <u>90.43</u>                   | 45min       |
| <b>Ours</b>                                              | <b>5114.22</b>      | <u><b>88.32</b></u>         | <b>89.89</b>                   | <b>30s</b>  |

geometric fidelity. As illustrated in Tab. 2, our method enables precise and camera-controllable generation with consistent geometric integrity.

**Qualitative Results** As presented in Fig. 3, further highlight the advantages of Deformable Gaussian Diffusion. Our method generates 4D scenes that are visually more appealing, with greater temporal coherence and more accurate preservation of object structure and motion details than baseline methods. For example, our generated videos exhibit smoother transitions and fewer artifacts in dynamic regions compared to SaV and Mosca. This visual superiority stems from our model’s direct prediction of deformable 3D Gaussians, which provides a rich and continuous representation of the scene’s evolution over time, effectively capturing complex dynamics from a single image input. The dynamic motion generation capabilities of AC3D (Bahmani et al., 2024) and CameraCtrl (He et al., 2024) are inherited from their underlying 2D video DiT priors, often resulting in videos with limited dynamism.

**Generation Controllability** Another key advantage of generating an **explicit** scene representation is the ability to ensure physical consistency through deterministic video “rendering from the input camera path”. We validate this by quantifying camera pose fidelity. As shown in Table 3, we compare our method against AC3D using the Relative Pose Error (RPE) metric (Sturm et al., 2012) on our evaluation dataset, demonstrating a significant improvement in pose accuracy.

#### 4.4 ABLATION AND ANALYSIS

**Effect of Deformation Gaussian Field** Fig. 4 illustrates the importance of the deformation Gaussian module. Without this module, the model struggles to differentiate between camera movement and the motion of foreground objects. This inability to properly combine 3D Gaussian splats from different timestamps leads to motion blur, spike artifacts, and a general degradation in image quality. By employing the deformation Gaussian field, our model effectively fuses reconstruction information from various moments, thereby achieving higher visual quality.

**Effect of Explicit Representation** Our adoption of an explicit 3D Gaussian Splatting representation offers several key advantages over implicit models, as detailed in Table 3. Firstly, it enables superior

Table 3: This comparison of the Average Relative Pose Error (RPE) highlights our method’s superior performance over the implicit model, demonstrating enhanced accuracy in translation and rotation.

| Method                            | Avg. RPE (Translation) ↓ | Avg. RPE (Rotation) ↓ | Novel View Synthesis | Depth Rasterization | Real-time Interaction |
|-----------------------------------|--------------------------|-----------------------|----------------------|---------------------|-----------------------|
| Implicit 3D Models                | 3.001                    | 0.810                 | ✓                    | ✗                   | ✗                     |
| Explicit 3D Representation (Ours) | <b>0.012</b>             | <b>0.008</b>          | ✓                    | ✓                   | ✓                     |

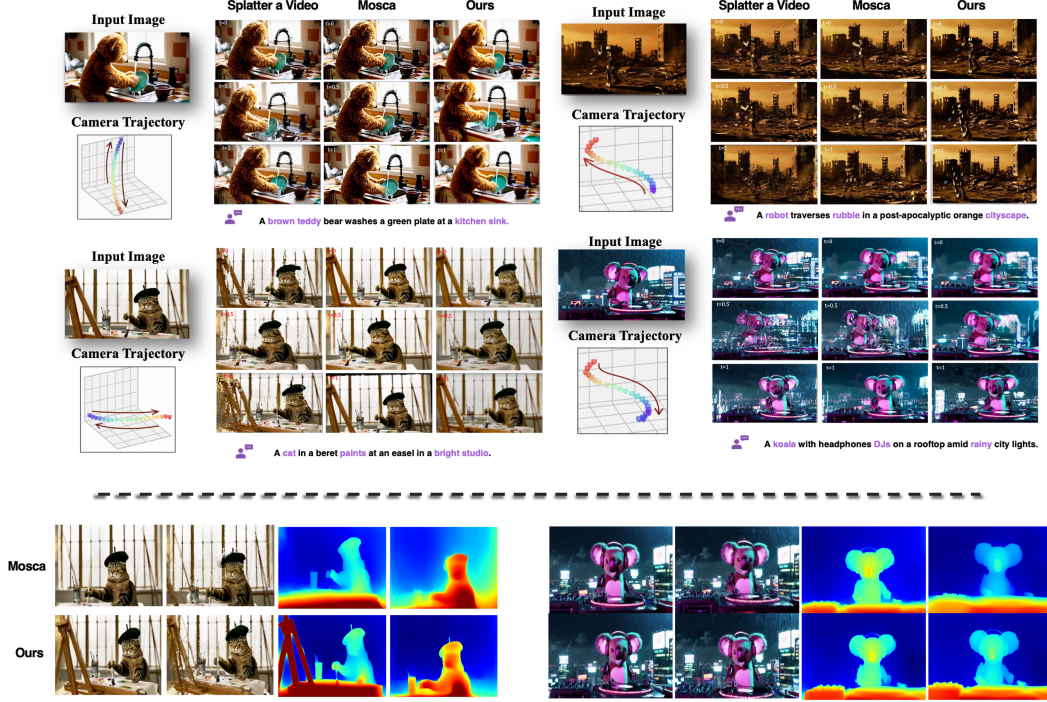


Figure 3: **Qualitative comparison with state-of-the-art methods.** DIFF4SPLAT (last column) generates more visually appealing and temporally consistent 4D scenes with superior geometric fidelity compared to baselines. Kindly zoom in for details.

camera controllability, drastically reducing the Relative Pose Error (RPE) in both translation and rotation. This ensures that the generated video faithfully adheres to the specified camera path. Secondly, the explicit nature of the representation unlocks additional functionalities not available in the implicit baseline, such as depth rasterization and real-time interaction. This not only enhances the model’s utility but also provides greater flexibility for downstream applications.

**Effect of Motion Loss** While photometric, geometric, and flow matching losses are prevalent techniques in 3D generation (Liang et al., 2024a), we conduct a detailed ablation study on the components of our proposed motion loss. The quantitative results, presented in Table 4, demonstrate its efficacy. Specifically, ablating the motion loss component prevents the network from accurately modeling temporal deformations, which is critical for dynamic video synthesis. The absence of this loss significantly degrades the quality of scene reconstruction and negatively impacts all quantitative evaluation metrics.

**Effect of Progressive Training** Direct dynamic training without static pretraining. We observe that omitting the static pretraining phase and directly engaging in dynamic training leads to a failure in the initialization of static 3DGS. This, in turn, results in unstable training dynamics and ultimately compromises the quality of the generated 4D scenes. Progressive training, starting with a static scene understanding, provides a robust foundation, ensuring stable 3DGS initialization and facilitating the subsequent learning of complex dynamic elements, thereby significantly enhancing the overall performance and visual fidelity. Direct dynamic training will converge to a suboptimal state or require significantly more training time (e.g., 21 days versus 7 days) for progressive training to reach a similar baseline quality. As illustrated in Figure 5, after 100K training iterations, our progressive



Table 4: **Ablation Study on Motion Loss.** We evaluate the impact of our proposed motion loss on dynamic video generation.

| Method          | FVD ↓          | KVD ↓        | QA-Quality ↑ | Avg. Matches ↑ | Subject Consistency Score ↑ | Background Consistency Score ↑ | Rec. Time ↓ |
|-----------------|----------------|--------------|--------------|----------------|-----------------------------|--------------------------------|-------------|
| w/o motion loss | 351.382        | 3.351        | 2.145        | 4821.56        | 82.45                       | 85.12                          | 30s         |
| Ours            | <b>210.153</b> | <b>2.316</b> | <b>2.813</b> | <b>5114.22</b> | <b>88.32</b>                | <b>89.89</b>                   | <b>30s</b>  |

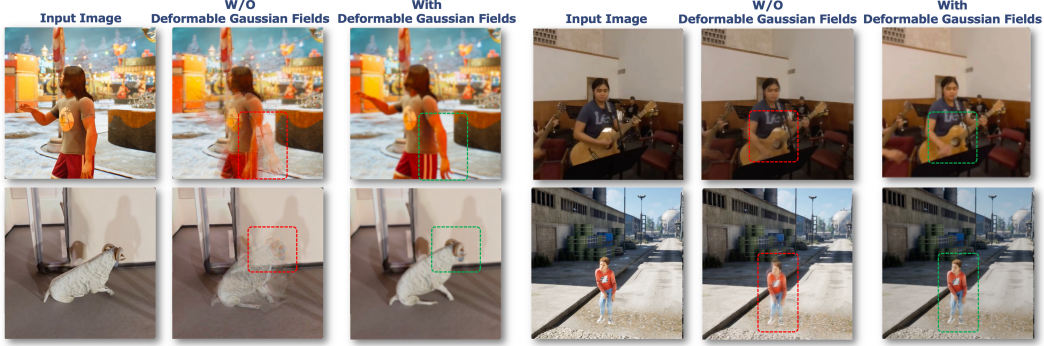


Figure 4: Ablation of the **Deformation Gaussian Field** shows that removing this module (the red bounding boxes) results in ghosting artifacts, particularly in the large motion frames.

training strategy yields significantly higher visual quality than direct dynamic training. This result underscores that progressive training not only enhances final performance and visual fidelity but also achieves superior results within the same computational budget, highlighting its resource efficiency.

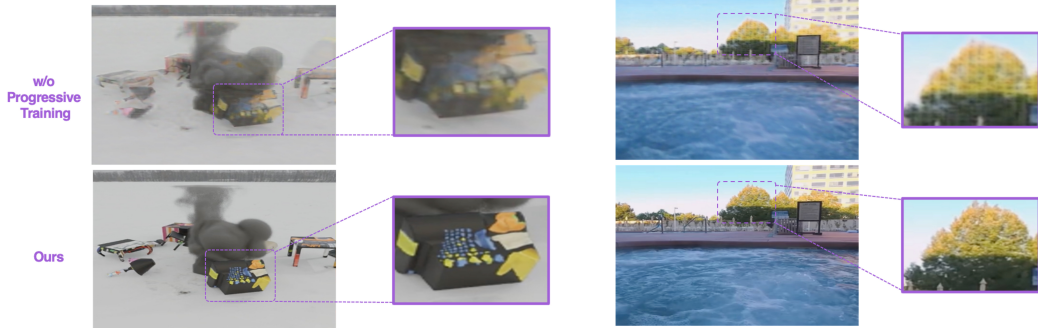


Figure 5: Ablation on the progressive training strategy.

## 5 CONCLUSION

In this work, we present a novel framework for explicit deformation Gaussian field generation from a single image in a *feed-forward* manner and achieves three key innovations: (1) unified diffusion transformer architecture integrating dynamic scene modeling, (2) geometry-aware latent representation enabling efficient view synthesis, (3) real-time rendering pipeline supporting practical applications. Extensive experiments demonstrate that our method achieves state-of-the-art performance in both geometric fidelity and computational efficiency, while eliminating the need for costly test-time optimization. We believe this work opens new opportunities for controllable 4D content creation at scale, bridging the gap between generative models and physically grounded scene understanding.

**Limitations and Future Work** While our method achieves superior performance and efficiency, video generation remains the computational bottleneck. This could be addressed through parallel inference or optimized denoising strategies. Future work will focus on extending temporal coherence modeling and material property prediction.

## REFERENCES

- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *arXiv preprint arXiv:2305.13245*, 2023. 4
- 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. *arXiv preprint arXiv:2411.18673*, 2024. 4, 6, 7
- Haotian Bai, Yuanhuiyi Lyu, Luta Jiang, Sijia Li, Haonan Lu, Xiaodong Lin, and Lin Wang. Comonerf: Text-guided multi-object compositional nerf with editable 3d scene layout. *arXiv preprint arXiv:2303.13843*, 2023a. 3
- Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. Recammaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025. 4, 17
- 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*, 2023b. 18
- 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*, 2023a. 2
- Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proc. CVPR*, 2023b. 4
- Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020. 4, 17
- Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. 4, 17
- Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025b. 2, 3
- Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 3
- Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *Advances in Neural Information Processing Systems*, 37: 96181–96206, 2025. 3
- Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023. 3
- Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017. 3
- Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *ICCV*, 2023. 2
- Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. GraphDreamer: Compositional 3D scene synthesis from scene graphs. *Proc. CVPR*, 2024. 3
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

- Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024. 3, 7
- 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, 2025. URL <https://arxiv.org/abs/2503.10592>. 3
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 2020. 3
- Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2
- Yan-Bin Jia. Plücker coordinates for lines in the space. *Problem Solver Techniques for Applied Computer Science, Com-S-477/577 Course Handout*, 2020. 3
- Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4d: Learning how things move in 3d from internet stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 4, 17
- Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4, 17
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 2023a. 3
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM TOG*, 2023b. 2
- Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2, 3, 6, 7, 20
- Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r. *arXiv:2406.09756*, 2024. 6
- Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024. 3
- Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *Proc. ICLR*, 2025a. 5
- Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023a. 4, 17
- Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023b. 2
- Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023c. 3

- Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast, and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. 4, 17
- 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. *arXiv preprint arXiv:2412.12091*, 2024a. 2, 3, 8
- 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, December 2024b. 4
- 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, December 2024c. 2
- Hanxue Liang, Jiawei Ren, Ashkan Mirzaei, Antonio Torralba, Ziwei Liu, Igor Gilitschenski, Sanja Fidler, Cengiz Oztireli, Huan Ling, Zan Gojcic, et al. Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. *arXiv preprint arXiv:2412.03526*, 2024d. 3
- Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. Gaufré: Gaussian deformation fields for real-time dynamic novel view synthesis. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2642–2652. IEEE, 2025. 4
- Chenguo Lin and Yadong Mu. Instructscene: Instruction-driven 3d indoor scene synthesis with semantic graph prior. *arXiv preprint arXiv:2402.04717*, 2024. 3
- Chenguo Lin, Yuchen Lin, Panwang Pan, Xuanyang Zhang, and Yadong Mu. Instructlay-out: Instruction-driven 2d and 3d layout synthesis with semantic graph prior. *arXiv preprint arXiv:2407.07580*, 2024a. 3
- Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. Diffsplat: Repurposing image diffusion models for scalable gaussian splat generation. *arXiv preprint arXiv:2501.16764*, 2025a. 3
- Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21136–21145, 2024b. 4
- Yuchen Lin, Chenguo Lin, Panwang Pan, Honglei Yan, Yiqiang Feng, Yadong Mu, and Katerina Fragkiadaki. Partcrafter: Structured 3d mesh generation via compositional latent diffusion transformers, 2025b. URL <https://arxiv.org/abs/2506.05573>. 3
- Yuchen Lin, Chenguo Lin, Jianjin Xu, and Yadong Mu. Omniphysgs: 3d constitutive gaussians for general physics-based dynamics generation. *arXiv preprint arXiv:2501.18982*, 2025c. 3
- Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. D13dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024. 3
- Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023. 5
- Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. Reconx: Reconstruct any scene from sparse views with video diffusion model. *arXiv preprint arXiv:2408.16767*, 2024. 3
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6



- Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4981–4991, 2023. 4, 17
- B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 3
- Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 2
- Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proc. CVPR*, 2024. 2
- Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiayi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*, 2023. 19
- Muyao Niu, Xiaodong Cun, Xintao Wang, Yong Zhang, Ying Shan, and Yinqiang Zheng. Mofa-video: Controllable image animation via generative motion field adaptations in frozen image-to-video diffusion model. *arXiv preprint arXiv:2405.20222*, 2024. 2
- Panwang Pan, Zhuo Su, Chenguo Lin, Zhen Fan, Yongjie Zhang, Zeming Li, Tingting Shen, Yadong Mu, and Yebin Liu. Humansplat: Generalizable single-image human gaussian splatting with structure priors, 2024. URL <https://arxiv.org/abs/2406.12459>. 4
- Elia Peruzzo, Vedit Goel, Dejia Xu, Xingqian Xu, Yifan Jiang, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Vase: Object-centric appearance and shape manipulation of real videos. *arXiv preprint arXiv:2401.02473*, 2024. 2
- Ryan Po and Gordon Wetzstein. Compositional 3D scene generation using locally conditioned diffusion. *Proc. 3DV*, 2024. 3
- 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 *International conference on machine learning*, pp. 8748–8763. PMLR, 2021. 6, 19
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. In *Proc. JMLR*, 2020. 6
- Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022a. 4
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022b. 3
- Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Computer Vision and Pattern Recognition (CVPR)*, 2016. 17
- Christoph Schuhmann. CLIP+MLP Aesthetic Score Predictor. <https://github.com/christophschuhmann/improved-aesthetic-predictor>, 2023. 6

- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 2022. 19
- QiuHong Shen, Xuanyu Yi, Mingbao Lin, Hanwang Zhang, Shuicheng Yan, and Xinchao Wang. Seeing world dynamics in a nutshell, 2025. URL <https://arxiv.org/abs/2502.03465>. 2, 3
- Vincent Sitzmann, Semon Rezhchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *Proc. NeurIPS*, 2021. 3
- Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IEEE/RSJ international conference on intelligent robots and systems*, 2012. 7
- 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*, 2024a. 3, 20
- Yang-Tian Sun, Yi-Hua Huang, Lin Ma, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Splatter a video: Video gaussian representation for versatile processing, 2024b. URL <https://arxiv.org/abs/2406.13870>. 7
- Yang-Tian Sun, Yihua Huang, Lin Ma, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Splatter a video: Video gaussian representation for versatile processing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024c. 3
- 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*, 2025. 3
- Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6, 19
- Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. *International Conference on Learning Representations (ICLR)*, 2019. 6
- Alexander Vilesov, Pradyumna Chari, and Achuta Kadambi. Cg3d: Compositional generation for text-to-3d via gaussian splatting. *arXiv preprint arXiv:2311.17907*, 2023. 3
- 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. *arXiv preprint arXiv:2412.04462*, 2024a. 3
- Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025a. URL <https://arxiv.org/abs/2503.11651>. 3
- Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b. 3, 7
- Wenshan Wang, DeLong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *International Conference on Intelligent Robots and Systems (IROS)*, 2020. 4, 17
- 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025c. 3

- Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *SIGGRAPH Conference*, 2024b. 2
- Christopher Wewer, Kevin Raj, Eddy Ilg, Bernt Schiele, and Jan Eric Lenssen. latentsplat: Autoencoding variational gaussians for fast generalizable 3d reconstruction. In *European Conference on Computer Vision*, pp. 456–473. Springer, 2024. 3
- Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20310–20320, 2024a. 3
- Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Chunyi Li, Liang Liao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtai Zhai, and Weisi Lin. Q-align: Teaching Imms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 6, 19
- 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. *arXiv preprint arXiv:2411.18613*, 2024b. 2, 3, 20
- 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*, 2024. 3
- Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023. 2
- Jinbo Xing, Hanyuan Liu, Menghan Xia, Yong Zhang, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Toonrafter: Generative cartoon interpolation. *arXiv preprint arXiv:2405.17933*, 2024a. 2
- Jinbo Xing, Menghan Xia, Yuxin Liu, Yuechen Zhang, Yong Zhang, Yingqing He, Hanyuan Liu, Haoxin Chen, Xiaodong Cun, Xintao Wang, et al. Make-your-video: Customized video generation using textual and structural guidance. *IEEE TVCG*, 2024b. 2
- Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024. 3
- Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass, 2025. URL <https://arxiv.org/abs/2501.13928>. 3
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024a. 6
- Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction, 2023. URL <https://arxiv.org/abs/2309.13101>. 2
- Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20331–20341, 2024b. 2, 4
- Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023. 3
- Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023. 2

- 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, 2024a. URL <https://arxiv.org/abs/2406.07472>. 3
- Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *arXiv preprint arXiv:2406.09394*, 2024b. 3
- 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*, 2024c. 3
- Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3
- Jinyan Yuan, Bangbang Yang, Keke Wang, Panwang Pan, Lin Ma, Xuehai Zhang, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. Immersegen: Agent-guided immersive world generation with alpha-textured proxies. *arXiv preprint arXiv:2506.14315*, 2025. 3
- Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 3
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5
- Weiwei Zhang, Jian Sun, and Xiaoou Tang. Cat head detection-how to effectively exploit shape and texture features. In *European conference on computer vision*, 2008. 19
- Xiaoming Zhao, Alex Colburn, Fangchang Ma, Miguel Angel Bautista, Joshua M. Susskind, and Alexander G. Schwing. Pseudo-generalized dynamic view synthesis from a video. In *International Conference on Learning Representations (ICLR)*, 2024a. 3
- 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*, 2024b. 3
- Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. VBench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025. 6
- Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4, 17
- Shijie Zhou, Zhiwen Fan, Dejia Xu, Haoran Chang, Pradyumna Chari, Tejas Bharadwaj, Suyu You, Zhangyang Wang, and Achuta Kadambi. Dreamscene360: Unconstrained text-to-3d scene generation with panoramic gaussian splatting. In *European Conference on Computer Vision (ECCV)*, 2024. 3
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyfe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018a. 3
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyfe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *Transactions on Graphics (TOG)*, 2018b. 4, 17
- Haoyi Zhu, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Chunhua Shen, Jiangmiao Pang, et al. Aether: Geometric-aware unified world modeling. *arXiv preprint arXiv:2503.18945*, 2025. 2, 3, 4



## A DECLARATION OF LLM USAGE

During the writing of the manuscript, we utilized a Large Language Model (ChatGPT) as a writing assistant. The scope of its use was limited to **improving grammar, polishing sentences, and enhancing the clarity and fluency of this manuscript**. The method, claims, experimental results and conclusions are developed by the authors.

## B DATASET CURATION.

As describe in **Section 3.1**, we construct a collection of 130,000 diverse videos featuring dynamic scenes captured by stationary cameras. Real-world datasets such as RealEstate10K only provide relative camera parameters estimated via COLMAP (Schonberger & Frahm, 2016), resulting in an unknown global scale. To address this, we re-estimate both metric depth maps and camera extrinsics using recent foundation models, Video Depth Anything (Chen et al., 2025a) and MegaSaM (Li et al., 2025b), to recover aligned geometry across frames.

---

### Algorithm 1 Metric Depth Reconstruction via Relative Depth Alignment

---

```

1: Input: RGB Image  $I$ , pre-trained DepthAnything model  $\mathcal{F}_{DA}$ , MegaSaM model  $\mathcal{F}_{MS}$ , metric
   depth oracle  $\mathcal{P}_M$ 
2: Output: Dense and metrically-scaled depth map  $D^*$ 

3:  $D_{rel} \leftarrow \mathcal{F}_{DA}(I)$  ▷ Generate relative depth map
4:  $\mathcal{S} \leftarrow \mathcal{F}_{MS}(I)$  ▷ Generate segmentation mask set
5:  $\mathcal{A} \leftarrow \emptyset$  ▷ Initialize anchor point set

6: for each mask  $M_i \in \mathcal{S}$  do
7:    $d_{gt,i} \leftarrow \mathcal{P}_M(M_i)$  ▷ Query ground-truth metric depth for the mask
8:   if  $d_{gt,i}$  is a valid measurement then
9:      $V_i \leftarrow \{D_{rel}(u, v) \mid M_i(u, v) = 1\}$  ▷ Extract corresponding relative depth values
10:     $d_{rel,i} \leftarrow \text{median}(V_i)$  ▷ Compute a robust representative value
11:     $\mathcal{A} \leftarrow \mathcal{A} \cup \{(d_{rel,i}, d_{gt,i})\}$  ▷ Add the pair to the anchor set
12:   end if
13: end for

14: ▷ Estimate optimal scale and shift by solving the least-squares problem
15:  $(s^*, t^*) \leftarrow \arg \min_{s, t} \sum_{(d_{rel,i}, d_{gt,i}) \in \mathcal{A}} (s \cdot d_{rel,i} + t - d_{gt,i})^2$ 

16: ▷ Apply the transformation to the full relative depth map
17:  $D^* \leftarrow s^* \cdot D_{rel} + t^*$ 

18: return  $D^*$ 

```

---

Table 5: **Training Datasets Statistics.** Overview of the datasets used for training **DIFF4SPLAT** at scale, highlighting their dynamic nature, multi-camera setups, depth annotations, tracking capabilities, and real-world applicability.

| Dataset                              | Dynamic? | Multi-camera? | Depth? | Tracking? | Real? | #Scenes | #Frames |
|--------------------------------------|----------|---------------|--------|-----------|-------|---------|---------|
| TartanAir (Wang et al., 2020)        | ✗        | ✗             | ✓      | ✗         | ✗     | 0.4K    | 0.49M   |
| MatrixCity (Li et al., 2023a)        | ✗        | ✗             | ✓      | ✗         | ✗     | 4.5K    | 0.31M   |
| RealEstate10K (Zhou et al., 2018b)   | ✗        | ✗             | ✗      | ✗         | ✓     | 70K     | 6.36M   |
| PointOdyssey (Zheng et al., 2023)    | ✓        | ✗             | ✓      | ✓         | ✗     | 0.1K    | 0.18M   |
| DynamicReplica (Karaev et al., 2023) | ✓        | ✗             | ✓      | ✓         | ✗     | 0.5K    | 0.26M   |
| Spring (Mehl et al., 2023)           | ✓        | ✗             | ✓      | ✗         | ✗     | 0.03K   | 0.003M  |
| VKITTI2 (Caban et al., 2020)         | ✓        | ✗             | ✓      | ✗         | ✗     | 0.1K    | 0.03M   |
| MultiCamVideo (Bai et al., 2025)     | ✓        | ✓             | ✗      | ✗         | ✗     | 14K     | 11M     |
| Stereo4D (Jin et al., 2025)          | ✓        | ✗             | ✓      | ✓         | ✓     | 74K     | 14.8M   |

## C MORE IMPLEMENTATION SETTINGS

**Reproducibility** To facilitate reproducibility, we present our detailed experimental settings and evaluation metrics in Section 4.1. This section provides a comprehensive description of our implementation details. Moreover, **our source code and pre-trained models will be publicly available.**

### C.1 VIDEO TRANSFORMER DENOSING DETAILS

**Details of Model Inputs** The model is conditioned on a single source image and a predefined camera motion trajectory, such as spiral, forward, backward, upward, or downward. Accompanying this, a textual prompt is provided, which can either be automatically generated from the source image using a Multimodal Large Language Model (MLLM) (Bai et al., 2023b) or set to a generic high-fidelity description, for instance, “a scene with 4K ultra HD, surround motion, realistic tone, panoramic shot, wide-angle view, cinematic quality”.

**Classifier-Free Guidance** Classifier-Free Guidance (CFG) has emerged as a prevalent technique for balancing controllability and sample diversity in diffusion models. However, we observe that its uniform scaling mechanism inadvertently introduces “over-sharpening artifacts” in the final frames of generated orbital sequences. To mitigate this limitation, we introduce a cosine-based dynamic guidance schedule during the sampling of validation videos, formulated as:

$$\gamma(t) = 1 + \gamma_{\max} \cdot \left( \frac{1 - \cos\left(\pi \left(\frac{N-t}{N}\right)^5\right)}{2} \right) \quad (6)$$

where  $\gamma_{\max}$  denotes the maximum guidance scale,  $N$  represents the total number of inference steps, and  $t$  is the current timestep. This adaptive scheduling progressively reduces guidance intensity in later denoising stages, effectively preserving temporal consistency while maintaining sample fidelity. In our experiments, we set the total number of inference steps  $N = 30$  and the maximum guidance scale  $\gamma_{\max} = 7.5$ .

### C.2 DEFORMATION FIELD GENERATION

To predict the per-Gaussian deformations, our **LDRM** employs a lightweight spatio-temporal network. The network takes as input a latent representation of the scene at a canonical time step, conditioned on a time embedding for the target frame  $t$ . The architecture extracts features at multiple spatial resolutions to effectively capture both local and global motion patterns. The final layer of the network is a convolutional layer with a kernel size of  $1 \times 1$ , which projects the high-dimensional features into the final deformation map  $\mathcal{D}$ . This map has a dimensionality of  $K_d = 10$  channels, which directly correspond to the predicted mean displacement (3 channels), rotational delta quaternion (4 channels), and scaling adjustment (3 channels) for each Gaussian primitive. No activation function is applied to the output layers for displacement and scale, allowing for unbounded predictions. The output quaternion components are normalized to ensure they represent a valid rotation.

### C.3 DETAILS OF PROGRESSIVE TRAINING SCHEME.

Our progressive training scheme’s efficacy in decoupling static and dynamic scene components is empirically validated. Initially, the model trains exclusively on static scenes, learning to predict an **identity deformation**. In this stage, positional and scaling offsets ( $\Delta\mu_p^t, \Delta s_p^t$ ) converge to zero, and rotational deformations ( $\Delta q_p^t$ ) approach the identity quaternion, yielding a static representation as canonical Gaussians remain untransformed. Dynamic scenes are introduced in a subsequent fine-tuning stage. This decoupling is enabled by our Gaussian deformation formulation:

$$\mu_p^t := \mu_p^0 + \Delta\mu_p^t, \quad q_p^t := q_p^0 \otimes \Delta q_p^t, \quad s_p^t := s_p^0 + \Delta s_p^t. \quad (7)$$

This design inherently separates the prediction of the canonical scene structure ( $\mu_p^0, q_p^0, s_p^0$ ) from its temporal evolution ( $\Delta\mu_p^t, \Delta q_p^t, \Delta s_p^t$ ).

Table 6: **Capability Comparison.** An explicit 4D representation enables a wide range of functionalities not supported by standard 2D video generation models.

| Capability                        | AC3D (Implicit 3D Models) | Ours (Explicit 4D Repr.)  |
|-----------------------------------|---------------------------|---------------------------|
| Novel View Synthesis              | ✓                         | ✓                         |
| Depth Rasterization               | ✗                         | ✓                         |
| Geometry Extraction               | ✗                         | ✓                         |
| Real-time Interaction             | ✗                         | ✓                         |
| Interactive exploration Latency ↓ | 28000 ms                  | <b>6.7 ms</b> (↓ 99.98%)  |
| Avg. Matches ↑                    | 2489.16                   | <b>5114.22</b> (↑ 105.5%) |
| Subject Consistency Score ↑       | 75.64                     | <b>88.32</b> (↑ 16.8%)    |
| Background Consistency Score ↑    | 75.91                     | <b>89.89</b> (↑ 18.4%)    |
| Cycle-Consistency ↑               | 20.68 dB                  | <b>34.5 dB</b> (↑ 66.8%)  |

#### C.4 DETAILS OF LOSS FUNCTION WEIGHTING

The loss weights ( $\lambda_p = 0.5$ ,  $\lambda_m = 2$ ) were determined empirically through a series of experiments on a validation set. We started with equal weights and adjusted them to ensure that the model did not prioritize one objective at the expense of others.

#### D EVALUATION PROTOCOL

To comprehensively evaluate our model, we utilize a suite of established metrics, Specifically:

❶ **Fréchet Video Distance (FVD) and Kernel Video Distance (KVD)** (Unterthiner et al., 2018):

These metrics evaluate the quality and temporal coherence of generated videos by measuring the distance between the feature distributions of real and generated video sets. Lower scores for both FVD and KVD indicate higher fidelity and better temporal consistency.

❷ **CLIP-Score** (Radford et al., 2021): This metric quantifies the semantic similarity between the generated video frames and the input text prompt. It leverages the joint text-image embedding space of the CLIP model, where higher scores signify better alignment between the generated content and the textual description.

❸ **CLIP-Aesthetic** (Schuhmann et al., 2022): We use a model built upon CLIP embeddings to predict the aesthetic quality of the generated content. This model is trained on datasets with human aesthetic ratings, and a higher score suggests a more visually pleasing result.

❹ **QA-Quality** (Wu et al., 2023): This refers to a Visual Question Answering (VQA)-based evaluation, where a LLaMA2-powered model is employed to assess the logical consistency and objective quality of the generated scenes. The model assigns a score on a range from 0 to 5, where a higher score indicates superior quality.

❺ **Temporal Consistency Metrics (Avg. Matches, Subject Cons. and Bg. Cons.):** Inspired by Video-bench (Ning et al., 2023), to specifically measure temporal stability, we use metrics based on dense optical flow or feature matching. Avg. Matches quantifies overall frame-to-frame consistency. Subject Consistency Score and Background Consistency Score measure the stability of the foreground subject and the background, respectively, after performing segmentation. Higher values for these metrics indicate smoother and more coherent videos.

#### E IMPLICIT VS. EXPLICIT 3D REPRESENTATIONS

Our work targets 4D scene generation by producing an “explicit” 3D representation (e.g., dynamic 3DGS), which offers capabilities substantially exceeding those of 2D video models. As demonstrated in Table 6, and inspired by prior work such as CAT4D (Zhang et al., 2008), an explicit 3D representation is a critical advantage for applications that demand a concrete understanding of and interaction with the world, including robotics and AR/VR.

Explicit 3D representations serve as a “**memory module**”, ensuring the consistency of the generated scenes. Unlike video generation models that predict 2D frames sequentially, our approach inherently enforces 3D consistency by predicting a single, unified explicit representation. Furthermore, 4D consistency is ensured by a training objective calculated from rendering the deformed 3D Gaussian representation from multiple viewpoints and at various timestamps. As shown in Table 6, we generate videos depicting a full 360-degree camera rotation. The resulting scenes exhibit seamless looping, where the final frame aligns perfectly with the first, showing no discernible seams or drift. We quantitatively verify this strong temporal consistency by measuring the similarity between the first and last frames (a.k.a., Cycle-Consistency), achieving a PSNR of 34.5 dB.

## F FEED-FORWARD VS. PER-SCENE OPTIMIZATION

Existing methods that produce explicit 3D outputs, rely on a time-consuming, post-hoc optimization process to reconstruct scenes from generated videos. For instance, DimensionX (Sun et al., 2024a) requires **1.3K GPU hours** to perform scene optimization from a single video. Even state-of-the-art 4D reconstruction algorithms like Mosca (Lei et al., 2024) require approximately **0.5 hours** to process one input video. The primary motivation of this work is therefore to unify these disparate stages into a single, efficient, feed-forward framework capable of generating a 4D representation in approximately **30 seconds**, achieving 60× acceleration. Our model is designed for efficiency and scalability, enabling dynamic scene reconstruction in a matter of seconds, which is a critical feature for many real-world applications where speed is essential.

Compared to per-scene optimization methods, our proposed approach achieves a substantial reduction in memory consumption during the reconstruction process, decreasing from 80GB to 25GB (a 3.2× reduction) in the same setting. This efficiency gain stems from the elimination of gradient computation requirements. Furthermore, we claim that the two approaches are not mutually exclusive. As explored in recent work like CAT4D (Wu et al., 2024b), efficient, end-to-end models can serve as an excellent initialization for optimization-based methods, significantly accelerating their convergence. This potential synergy further highlights that developing fast, feed-forward models is a valuable research direction.

In summary, considering both the reconstruction and rendering stages (e.g., maximum GPU memory), our approach remains competitive in terms of memory consumption compared to per-scene optimization methods.

## G FAILURE CASES

As shown in Figure 6, our method can produce artifacts when rendering novel timestamps, especially from disparate viewpoints. This issue, common to related methods, stems from ambiguity in estimating temporal deformations when propagating 3D Gaussians from multiple reference frames.

**Motion Ambiguity.** Single-image-to-4D generation is an ill-posed problem, as one image can imply multiple plausible motions (e.g., a bird gliding vs. flapping). This ambiguity can lead to inaccuracies in the predicted deformation field and corresponding visual artifacts. Incorporating more explicit motion priors in future work could address this limitation.

**Out-of-Distribution Generalization.** Model performance may degrade on out-of-distribution inputs, such as novel object categories or abstract artistic styles, resulting in lower-quality geometry and motion. Exploring few-shot domain adaptation techniques presents a promising direction for enhancing model robustness.

## H MORE VISUAL RESULTS

We provide more visualization results of **DIFF4SPLAT** in Figure 7, Figure 8, and Figure 9.

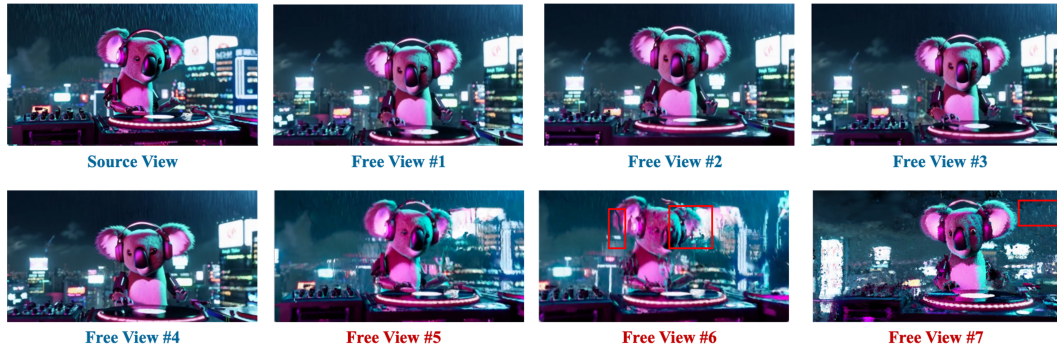


Figure 6: Failure Case. **DIFF4SPLAT** can produce artifacts when rendering novel timestamps, especially from disparate viewpoints. This issue, common to related methods, stems from ambiguity in estimating temporal deformations when propagating 3D Gaussians from multiple reference frames.

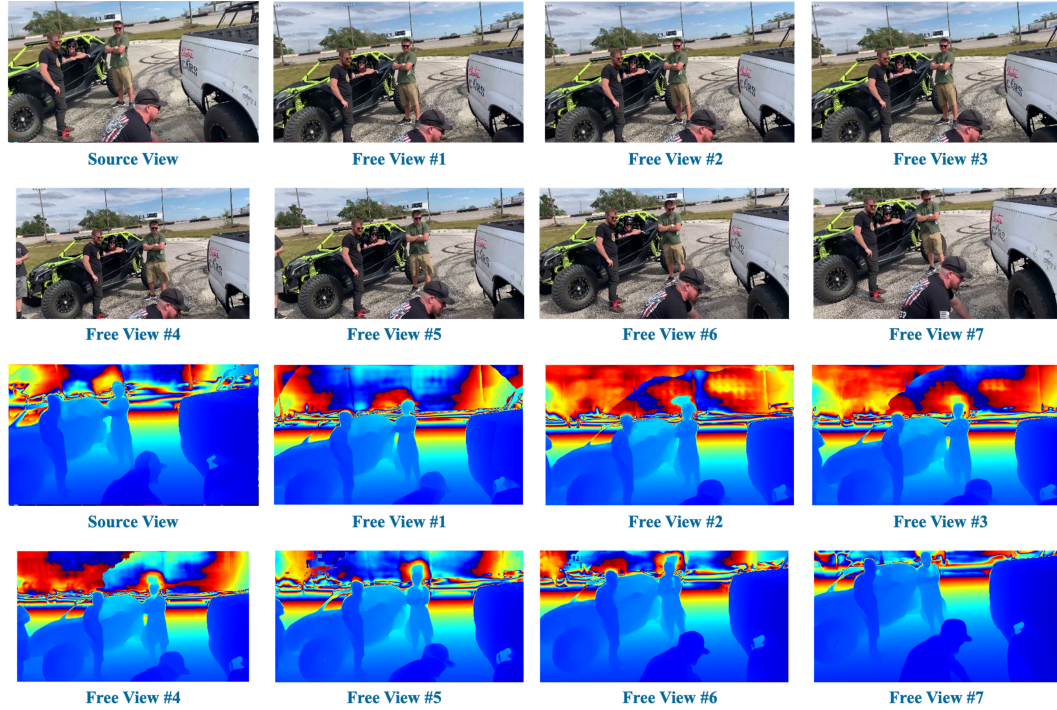
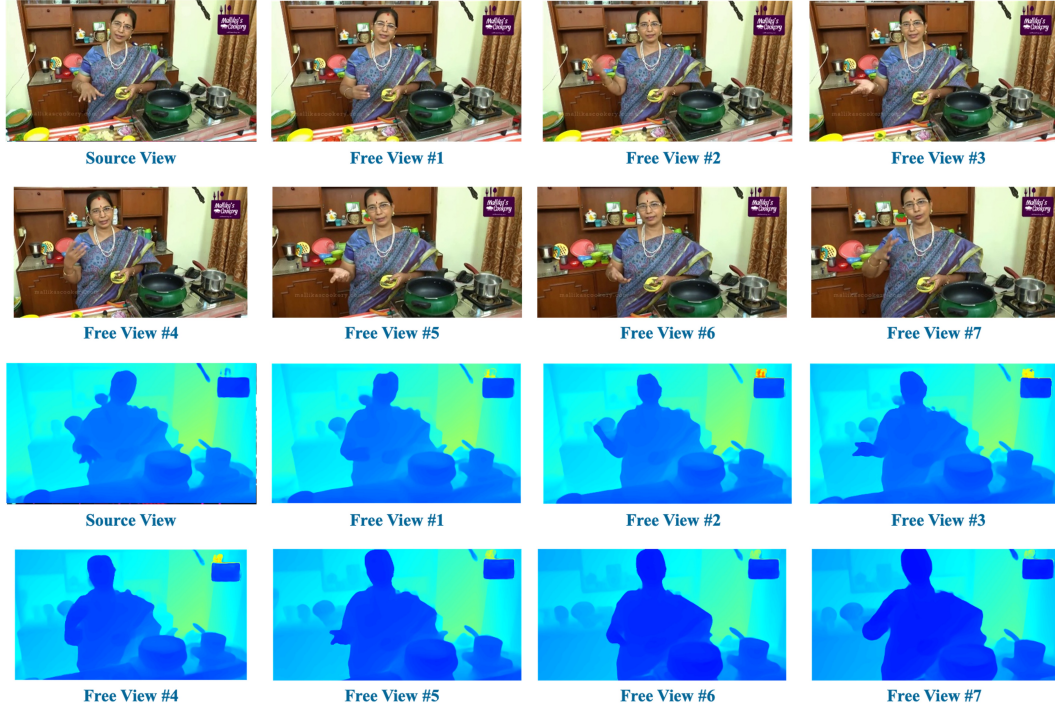
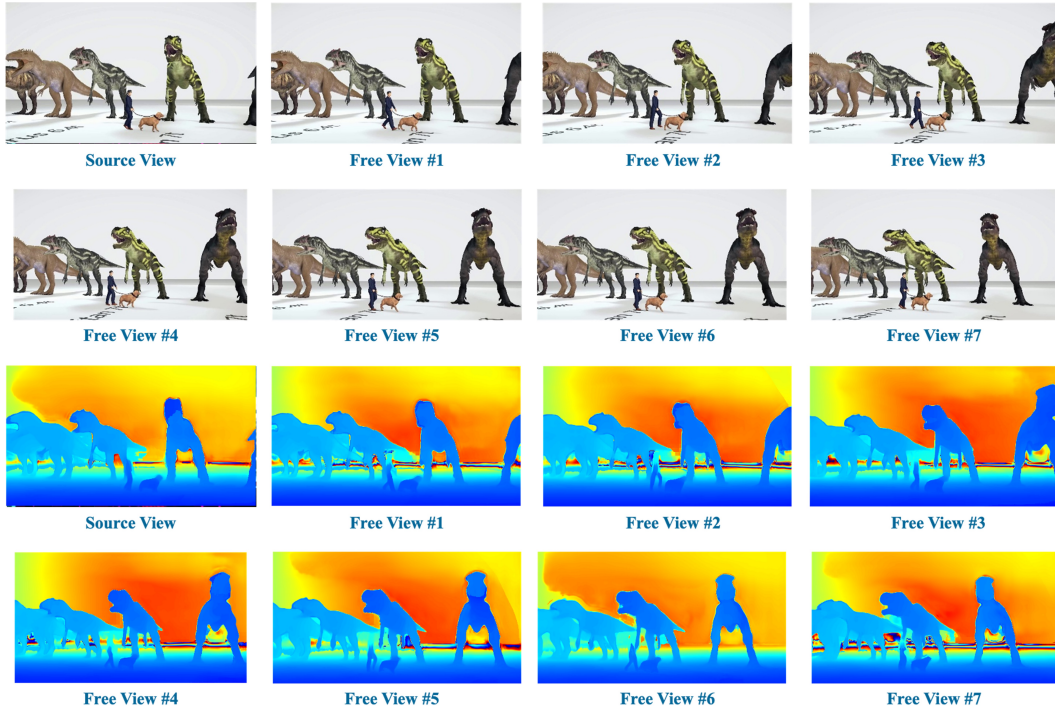


Figure 7: More qualitative of **DIFF4SPLAT** for 4D Scene generation.



Figure 8: More qualitative of **DIFF4SPLAT** for 4D Scene generation.Figure 9: More qualitative of **DIFF4SPLAT** for 4D Scene generation.