Geometry-Aware Single-Image 4D Synthesis via Dense Trajectory Generation

Yanran Zhang*,1, Ziyi Wang*,1, Wenzhao Zheng†,1, Zheng Zhu[2], Jie Zhou[1], and Jiwen Lu✉,1

[1] Department of Automation, Tsinghua University, China [2] GigaAI
https://ivg-yanranzhang.github.io/MoGe4D/
zhangyr21@mails.tsinghua.edu.cn, lujiwen@tsinghua.edu.cn

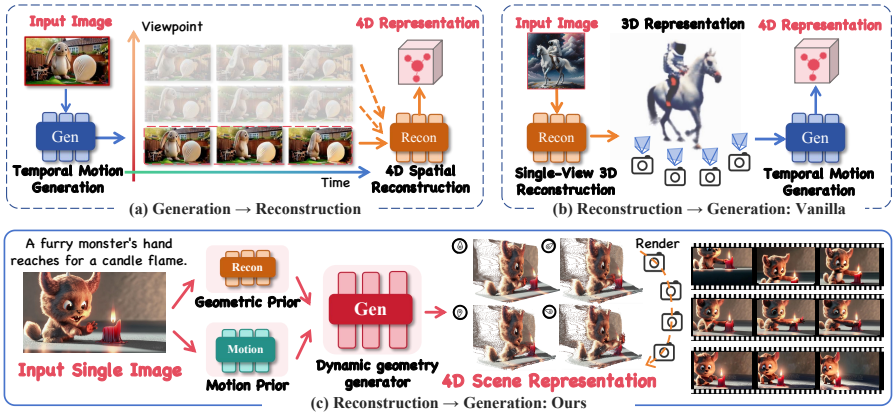

Fig. 1: MoGe4D for 4D synthesis from a single image. Most existing paradigms either suffer from geometric inconsistencies (*generate-then-reconstruct*) or are constrained by animating a pre-determined static geometry (vanilla *reconstruct-then-generate*). Our MoGe4D advances by tightly coupling geometric modeling and motion generation, effectively achieving consistent 4D motion and geometry.

Abstract. Generating interactive and dynamic 4D scenes from *a single static image* remains a core challenge. Most existing *generate-then-reconstruct* and *reconstruct-then-generate* methods decouple geometry from motion, causing spatiotemporal inconsistencies and poor generalization. To address these, we present **MoGe4D** (**Mo**tion and **Ge**ometry-Aware image-to-**4D** Synthesis), a geometry-conditioned framework for single-image 4D synthesis that models a scene as dense 4D point trajectories. Instead of treating geometry and dynamics as two disconnected stages, our method starts from an initial geometric prior inferred from the input image and predicts future time-varying trajectories in a diffusion process, improving spatiotemporal coherence while preserving structural stability. To support this task, we first introduce TrajScene-60K, a large-scale dataset of 60,000 video samples with dense 4D point trajectories, addressing the scarcity of high-quality training data for scene-

* Equal contributions. †Project leader. ✉Corresponding author.

level 4D generation. Built on this, our diffusion-based 4D Scene Trajectory Generator (4D-STraG) predicts geometry-consistent and motion-plausible trajectory fields conditioned on the input image, with a depth-guided motion normalization strategy to reduce scale ambiguity and a Motion Perception Module (MPM) to inject motion-aware priors. We further propose a 4D View Synthesis Module (4D-ViSM) to render the generated 4D representation into videos under arbitrary camera trajectories. Experiments show that MoGe4D produces high-quality 4D scenes with strong temporal coherence, favorable geometry-aware consistency, and compelling novel-view synthesis from a single image. Code: https://github.com/Zhangyr2022/MoGe4D.

Keywords: 4D Generation · Dense Trajectories · Geometry-Aware · Novel View Synthesis · Diffusion Models

1 Introduction

4D scene generation seeks to reconstruct comprehensive spatiotemporal representations capturing both explicit 3D geometry and complex temporal dynamics. Generating such dynamically rich 4D scenes from *a single static image* remains a fundamental challenge [41], as the model must infer plausible future motion from a static observation while preserving structural coherence across time and viewpoints. Achieving this capability would largely benefit applications such as virtual reality, augmented reality [19, 42], and immersive content creation.

Although recent video generation models can produce realistic dynamic content, they generally lack an explicit understanding of 3D structure, leading to view inconsistencies and failing to capture physically plausible motion. Existing single-image-to-4D methods therefore typically adopt one of two decoupled paradigms. *Generate-then-reconstruct* methods [37, 51, 63, 72] use powerful video models to synthesize multi-view videos before reconstructing a 4D representation. This paradigm benefits from the high-fidelity and rich dynamics offered by existing generation models. However, video models struggle to maintain strict geometric consistency across the generated views, leading to significant artifacts and structural collapse during the subsequent 3D reconstruction phase. Consequently, the alternative paradigm, *reconstruct-then-generate* [8, 24, 27, 48, 76], has emerged as another promising direction. By first establishing a static 3D structure, this approach provides a robust geometric foundation for subsequent motion generation. Still, by decoupling static geometry from motion generation, they discard the rich dynamic potential latent in the source image. As a result, they are restricted to modeling physically plausible and externally constrained motions (e.g., swinging) and struggle to generate large-scale and self-initiated movements that originate from the scene or objects themselves.

To address these challenges, we revisit the *reconstruct-then-generate* paradigm and make its interface between structure and dynamics substantially tighter, as illustrated in Figure 1. Rather than explicitly refining the base geometry during generation, we condition future dynamic generation on an initial geometric

prior inferred from the input image, and model the scene as dense 4D point trajectories, which is directly predicted from the single input image with an integrated model. To facilitate training of the 3D motion generation model, we first built TrajScene-60K, a large-scale dataset of 60,000 samples with 4D point trajectories, tackling data scarcity of large-scale, high-quality 4D scene data with complex dynamics. We then introduce the *4D Scene Trajectory Generator* (4D-STraG) that takes an input image and predicts future time-varying point trajectories relative to the initial frame. Unlike prior works that treat these as separate steps, 4D-STraG operationalizes them within a unified denoising process, producing coherent 4D point trajectories whose motion is intrinsically consistent with the evolving 3D structure. To fully leverage priors from the input image, we further incorporate a depth-guided motion normalization method to enhance geometric awareness, and a Motion Perception Module (MPM) to leverage plausible motion priors. Finally, the generated 4D point cloud trajectories are rendered into high-fidelity dynamic videos from arbitrary novel viewpoints using our *4D View Synthesis Module* (4D-ViSM), completing the full pipeline from a single image to a coherent 4D scene.

Experimental results of both quantitative and qualitative comparisons demonstrate that our method consistently outperforms existing baselines, producing 4D scenes with more pronounced dynamics, stronger 3D motion consistency, and superior performance. Detailed ablation studies further confirm the effectiveness of our key modules, highlighting how each contributes to coherent and physically plausible 4D motion. Our contributions can be summarized as:

- We construct TrajScene-60K, a large-scale 4D scene dataset featuring dense 4D point cloud trajectories, videos, and text to advance research in this field.
- We propose a geometry-conditioned dense-trajectory formulation for single-image 4D synthesis. It predicts scene trajectories from an initial geometric prior, avoiding the reliance on loosely decoupled stages.
- We develop 4D-STraG, a diffusion-based trajectory generator with depth-guided motion normalization and a Motion Perception Module for structural stability and motion plausibility, alongside 4D-ViSM for high-quality novel-view dynamic rendering from the generated 4D representation.
- Extensive experiments show that MoGe4D delivers strong perceptual quality, geometry-aware consistency, and efficient 4D generation.

2 Related Work

Video Generation. The field aims to create temporally coherent and visually realistic dynamic visual content. From early VAE [3,4,15,20,55] and GAN [13,54,56] frameworks to modern large-scale diffusion models [21,31,50,52,57] trained on extensive video data, the realism and resolution of generated videos have been revolutionized. In addition to text control generation, researchers have introduced various guidance strategies for video generation, such as structure-based [12,39,68], image-based [10,14], and temporal controls [49,62,69]. How-

ever, most of them are fundamentally limited as they operate in 2D pixel space, lacking explicit 3D scene modeling.

Novel-View Synthesis (NVS). 3D reconstruction-based methods reconstruct a 3D or 4D representation [9, 61, 77, 78], ensuring strong geometric consistency but often requiring costly optimization and suffering from artifacts. Generation-based methods use pretrained video diffusion models conditioned on camera trajectories to synthesize novel views [5, 17, 36, 43, 73]. While leveraging strong generative priors, they often exhibit inconsistencies and object drift under large viewpoint changes or complex scenes.

4D Generation. 4D generation aims to produce temporally evolving 3D representations from texts or sparse images. A growing body of work addresses 3D point tracking [16,28,29,32,33,44,65,66] and video-to-4D reconstruction [8,25,26, 59, 60, 63, 64, 67, 70, 71], estimating dense spatiotemporal correspondences from multi-frame video inputs—a setting where rich temporal signals substantially constrain the problem. Image-to-4D is more ill-posed than video-to-4D, requiring full 3D and temporal reconstruction from single images without multi-frame priors. Existing Image-to-4D methods largely follow two decoupled paradigms: *generate-then-reconstruct*, which first synthesizes videos and then reconstructs 4D (e.g., L4GM [47], 4Real [72], DimensionX [51], Free4D [37]), and *reconstruct-then-generate*, which builds static 3D assets before animating them (e.g., Animate124 [76], Animate3D [24], 4D-fy [4], Gen3C [48]). While effective, the former suffers from video–geometry mismatch, and the latter is often object-centric and limited in scene complexity. Recently, 4DNeX [11] and One4D [40] concatenates RGB-XYZ, in contrast to our method, which advances the reconstruct-then-generate paradigm by modeling geometry-aware dense trajectory generation and enables consistent 4D synthesis.

3 Dataset Curation – TrajScene-60K

Developing a robust 4D generative framework necessitates a multi-faceted dataset comprising three essential modalities: dense 4D point trajectories, viewpoint-specific visual observations, and high-level semantic descriptions of 4D environments. To address the acute shortage of high-quality, large-scale annotated data—especially for complex scene-level videos featuring intricate motion and non-rigid dynamics—we present **TrajScene-60K**. This meticulously curated dataset provides the foundational supervision required for learning consistent 4D scene representations and trajectories.

As shown in Figure 2, we construct our dataset from the high-quality WebVid-10M corpus [7], extracting about 200,000 candidates via a two-stage automated filtering pipeline. CogVLM2 [22] generates a detailed caption $\mathcal{C}$ for each video that captures both its visual content and the temporal characteristics of its motion. And DeepSeek-V3 [35] evaluates these captions against two criteria: (1) the presence of one or more clearly countable entities, and (2) the exhibition of self-initiated, non-rigid, or articulated motion. Videos dominated by unstructured dynamics (e.g., crowd behaviors, wind-driven oscillations, or background

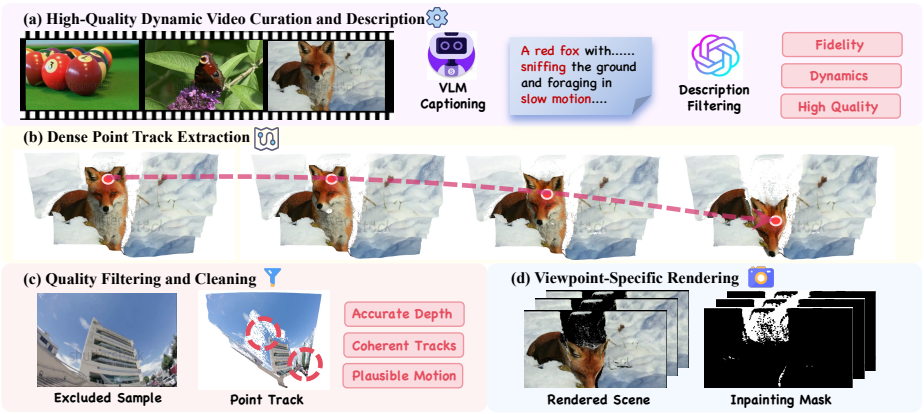

Fig. 2: TrajScene-60K curation pipeline. We curate videos from WebVid-10M, filtered via VLMs for structured motion and countable entities. Dense 4D point tracks are extracted and refined via depth filtering and Gaussian Splatting, producing 60K high-quality 4D scenes.

jitter) or by camera motion are discarded. This is particularly important for 4D learning, since egomotion would otherwise be conflated with genuine object dynamics and introduce spurious global drift into the pseudo-GT trajectories. Retaining only clips whose motion originates from the scene itself thus yields well-structured, quantifiable motion targets that enable learning of clear and interpretable dynamic representations.

To extract 4D dense point track data from videos, we employ the DELTA [44] model. The model takes RGB video sequences $\mathcal{V} \in \mathbb{R}^{T\times H\times W\times 3}$ as input and utilizes monocular depth estimation methods to obtain depth maps $\mathcal{D} \in \mathbb{R}^{T\times H\times W}$, subsequently estimating occlusion-aware 4D trajectories $\mathcal{P} \in \mathbb{R}^{T\times H\times W\times 4}$. Each 4D vector $\mathbf{p}_{t,u,v} = (u_t, v_t, d_t, o_t)$ represents the 3D position and occlusion status in frame t corresponding to the pixel located at (u, v) in the first frame. After obtaining 4D point cloud scene data, we apply a strict three-criterion quality filtering process: (1) samples containing trajectories with invalid or anomalous depth values (e.g., near-infinite or zero) are removed; (2) samples exhibiting excessively large standard deviation in scene depth, indicative of global depth estimation errors, are discarded; (3) a scale consistency check removes geometrically inconsistent samples by verifying that uniformly scaled point clouds yield identical renderings from the original camera perspective. This pipeline yields 60,000 high-quality samples at 596×336 resolution.

We then render these 4D scenes into videos from the original camera viewpoint using Gaussian Splatting [30], with inpainting masks generated for void regions. Compared to existing benchmarks, TrajScene-60K provides over **3 million frames** and approximately **12 billion 3D point annotations** with dense occlusion-aware tracking, per-frame depth, and language descriptions across diverse real-world indoor and outdoor environments—a combination of scale, realism, and multi-modal annotation unmatched by prior datasets. A detailed comparison with existing datasets is provided in the supplementary material.

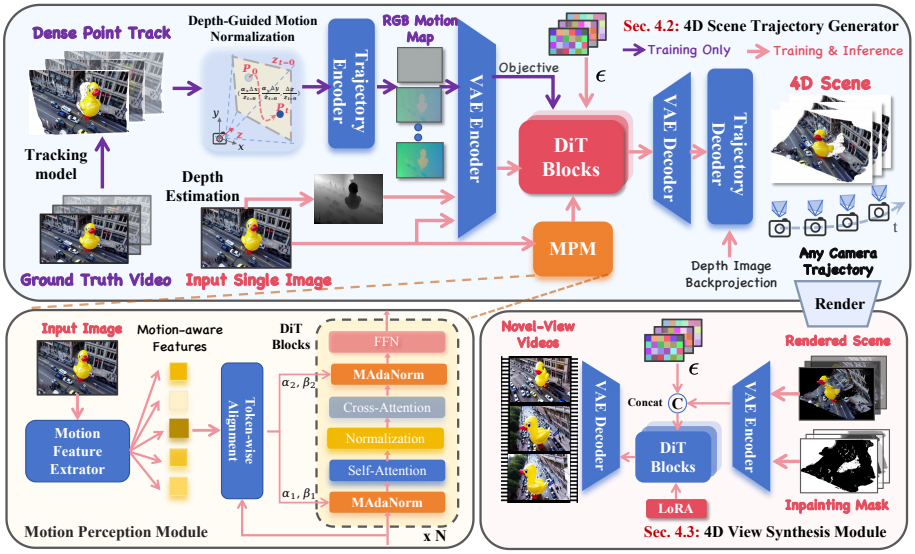

Fig. 3: Pipeline of MoGe4D. Top: The 4D Scene Trajectory Generator (Sec. 4.2), a Diffusion Transformer, jointly generates future geometry and motion. Bottom-Left: The Motion Perception Module (MPM) identifies potential motion regions and semantic structure from the input image. Bottom-Right: The 4D View Synthesis Module (Sec. 4.3) renders the output into novel-view videos.

4 Proposed Approach

4.1 Problem Definition

Given a single input image $\mathcal{I} \in \mathbb{R}^{H\times W\times 3}$ and its description $\mathcal{C}$, our goal is to generate a physically plausible and temporally consistent 4D scene. We represent the scene as a point cloud sequence $\mathcal{P} \in \mathbb{R}^{T\times N\times 3}$, capturing the 3D geometry and motion trajectories of $N = H \times W$ points over T frames. From this representation, we render dynamic scene videos $\mathcal{V} \in \mathbb{R}^{T\times H'\times W'\times 3}$ from arbitrary novel viewpoints or along arbitrary camera paths to the gap between a static image and full multi-view 4D dynamic generation.

Unlike canonical-mesh or template-based 4D representations, dense point trajectories are class-agnostic and require no predefined structure, decoupling geometry from appearance to flexibly accommodate diverse topologies and large deformations. This factorization provides a stable spatial reference and avoids the brittleness of loosely cascaded pipelines, while still allowing the model to synthesize rich scene dynamics in a unified trajectory space. To systematically achieve these objectives, we first design **4D S**cene **Tra**jectory **G**enerator (**4D-STraG**), a diffusion model that predicts future 4D trajectory fields conditioned on the initial geometry from the initial image. Then we present in detail the **4D Vi**ew **S**ynthesis **M**odule (**4D-ViSM**), which leverages the reconstructed 4D

representation to enable high-quality video generation under arbitrary camera motion. The overview illustration is shown in Figure 3.

4.2 4D Scene Trajectory Generator (4D-STraG)

To leverage the rich motion priors and structural understanding capabilities, we build upon Wan2.1 [57] and finetune its VAE and DiT modules separately.

Point Trajectory Initialization and Normalization. To enhance training stability and ensure compatibility with generative model scales, we only use the 4D-STraG to predict relative motion $\mathbf{\Delta P}_t = \mathbf{P}_t - \mathbf{P}_0 = \{[\Delta x_t, \Delta y_t, \Delta z_t]\}$, where $t \in [0, T]$ and $\mathbf{P}_0$ denotes coordinates in the first frame. We further normalize $\mathbf{\Delta P}_t$ to avoid an unlimited data value range. Based on the observation that a small 3D movement in a nearby object causes a large displacement on the 2D image plane, whereas the same movement in a distant object appears minuscule, we propose a *Depth-Guided Motion Normalization Strategy*. It normalizes the absolute motion of each point relative to its viewing frustum at the initial depth. By doing so, we transform raw motion into a scale-invariant representation, ensuring perceptual consistency across different distances and producing a more uniform data distribution to learn from.

Specifically, given focal lengths f_x, f_y and image size $W \times H$, we define the scaling factors $\alpha_x = f_x/W$ and $\alpha_y = f_y/H$. Geometrically, z/α_x and z/α_y correspond to the width and height of the viewing frustum at depth z, respectively. We normalize motion quantities by the viewing frustum size at depth $z = \mathbf{P}_0^{(z)}$:

$$\Delta\tilde{x}_t = \frac{\alpha_x \cdot \Delta x_t}{z}, \quad \Delta\tilde{y}_t = \frac{\alpha_y \cdot \Delta y_t}{z}, \quad \Delta\tilde{z}_t = \frac{\Delta z_t}{z}. \tag{1}$$

This depth-dependent normalization achieves scale invariance across different depth ranges, enabling our diffusion model to effectively learn motion patterns without being biased by the absolute spatial position of points. During inference, we use UniDepthv2 [46] to estimate depth, ensuring consistency with the DELTA tracking model setup. The relative motion maps are de-normalized and fused with the initial point cloud to form a 4D scene representation.

Model Pipeline. During training, our model takes an image–caption pair as input and learns a diffusion model to predict pixel-level point trajectories across frames. To adapt the generative backbone, we first finetune a motion-sensitive VAE capable of handling trajectory signals. Specifically, the relative point displacement $\mathbf{\Delta P_t}$ is transformed into an RGB motion map by a lightweight Trajectory Encoder, where spatial movements are represented as color variations while the shape and appearance remain static as the first frame. Correspondingly, a Trajectory Decoder is appended after the VAE decoder, ensuring accurate recovery of point trajectories from the RGB motion map.

After finetuning the VAE, we adapt the Diffusion Transformer (DiT) to handle latents encoded from the RGB motion map. We explicitly inject strong geometric priors into the model, thereby enhancing its generative capability. Specifically, the depth information from the initial frame is encoded into a latent representation using the VAE encoder. This provides the model with robust

structural priors and geometric cues, significantly improving its ability to reason about scene layout and object relationships. The image, noise, and depth latents are concatenated along the feature dimension to form the final input for the DiT:

$$z_{\text{combined}} = \text{Concat}(z_{\text{image}}, z_{\text{noise}}, z_{\text{depth}}). \tag{2}$$

The DiT model is trained using flow matching [34], which learns deterministic flows from noise to data distributions by minimizing the error between predicted and true flow fields, enabling accurate modeling of pixel-level motion. The objective function is defined as:

$$\mathcal{L}_{\text{fm}} = \mathbb{E}_{t,x_0,x_1}\left[|v_\theta(t, x_t) - (x_1 - x_0)|^2\right], \tag{3}$$

where x_0 and x_1 represent the initial and target states, respectively, x_t denotes the interpolated state at time t, and v_θ is the flow vector predicted by the model.

Motion Perception Module (MPM). Adapting powerful motion priors from image-to-video diffusion models poses a key challenge. These models excel at generating dynamic, moving scenes from a static image. Our objective, however, is to leverage this pretrained temporal knowledge for a distinct task: generating structurally static videos that exhibit dynamic color variations. Here, the 'dynamic color' sequence is an intermediate RGB motion-map representation that encodes per-pixel 3D displacements. This creates a core conflict, as the model must learn to translate its ingrained understanding of physical displacement into a new domain of temporal color evolution.

To resolve this conflict and effectively guide the model's adaptation, we introduce the Motion Perception Module (MPM). The MPM is designed to identify semantic regions within the static scene that are plausible candidates for these color dynamics. We first employ a pretrained motion feature extractor, OmniMAE [18], to derive motion-aware patch-level features $\mathbf{S}$ from the input static image. To embed motion information into the diffusion process, we introduce *Motion-aware Adaptive Normalization* (MAdaNorm). While inspired by conditional injection mechanisms like AdaLN [45], our proposed module is specifically designed for 4D motion generation. Unlike global conditioning approaches, MAdaNorm performs token-wise modulation driven by motion-aware features, enabling fine-grained control over temporal coherence in a geometrically consistent manner. Specifically, after spatial alignment of motion features $\mathbf{S}$ by resizing them to match the DiT token sequence length, token-wise adaptive parameters are generated through linear layers. For intermediate features $\mathbf{F}_t^i \in \mathbb{R}^{N\times d}$ in the i-th DiT block, the process is as follows:

$$\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 = \text{Linear}(\mathbf{S}), \tag{4}$$

$$\mathbf{F}' = \text{Attn}\left(\boldsymbol{\gamma_1}\boldsymbol{\alpha}_1 \odot \text{LN}(\mathbf{F}_t^i) + \boldsymbol{\gamma_1}\boldsymbol{\beta}_1\right), \tag{5}$$

$$\mathbf{F}'' = \text{MLP}\left(\boldsymbol{\gamma_2}\boldsymbol{\alpha}_2 \odot \text{LN}(\mathbf{F}') + \boldsymbol{\gamma_2}\boldsymbol{\beta}_2\right), \tag{6}$$

where $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^{N\times d}$ are token-wise scaling and bias parameters, $\boldsymbol{\gamma_1}, \boldsymbol{\gamma_2} \in \mathbb{R}^d$ are learnable global gating coefficient, and $\odot$ denotes token-wise multiplication. Experiments show that MPM improves the fidelity of motion trajectory modeling and enhances spatio-temporal consistency of generated 4D content.

4.3 4D View Synthesis Module (4D-ViSM)

After obtaining a dense 4D point cloud representation, we propose the 4D-ViSM to achieve novel view video synthesis along arbitrary camera trajectories. The original point cloud may not fully cover image regions in novel views, which results in holes in the rendered output. Concretely, we rasterize each frame of the time-varying point cloud as a per-frame set of 3D Gaussians; temporally linking these per-frame Gaussians yields a lightweight dynamic renderer that avoids optimizing a full dynamic-4DGS field while substantially reducing projection-induced holes. We thereby leverage generative models to complete these missing regions, ensuring visual coherence and plausibility.

Considering the excellent performance of Wan2.1 [57] in video generation tasks, we also choose to finetune this model to build our 4D-ViSM. The training data also come from our TrajScene-60K dataset, including rendered videos, corresponding occlusion masks, ground truth videos, and captions. During training, we follow the Wan2.1 mask processing strategy, setting the mask value to 0.5 for regions without projected points in each frame. Through finetuning, our model achieves high-quality and visually consistent novel view video synthesis.

4.4 Inference

Given a single image and a text prompt, we first estimate the initial depth with UniDepthv2 [46], matching the estimator used by the DELTA tracker during training. The VAE then encodes the image and depth into latents while MPM extracts motion features, after which the DiT generates and decodes relative motion latents. These are de-normalized by reversing our depth-guided motion normalization and fused with the initial point cloud to construct the 4D scene, which 4D-ViSM finally renders from arbitrary camera poses into a spatio-temporally consistent novel-view video.

5 Experiments

5.1 Experimental Setup

Implementation Details. Trained on TrajScene-60K dataset, our model generates 512×368 videos with a length of 49 frames. For the 4D-STraG module, we performed full-parameter training based on Wan2.1-14B [1, 57]. The trajectory encoder and decoder are both shallow ResNets. We first trained its tracking components and VAE decoder for 5k steps, then its DiT for 2k steps using OmniMAE features for motion conditioning. The 4D-ViSM module finetuned Wan2.1-14B [2, 57] with LoRA for 10k steps. We used AdamW [38] with a learning rate of 2×10^{-5}. All experiments ran on four NVIDIA H20 GPUs.

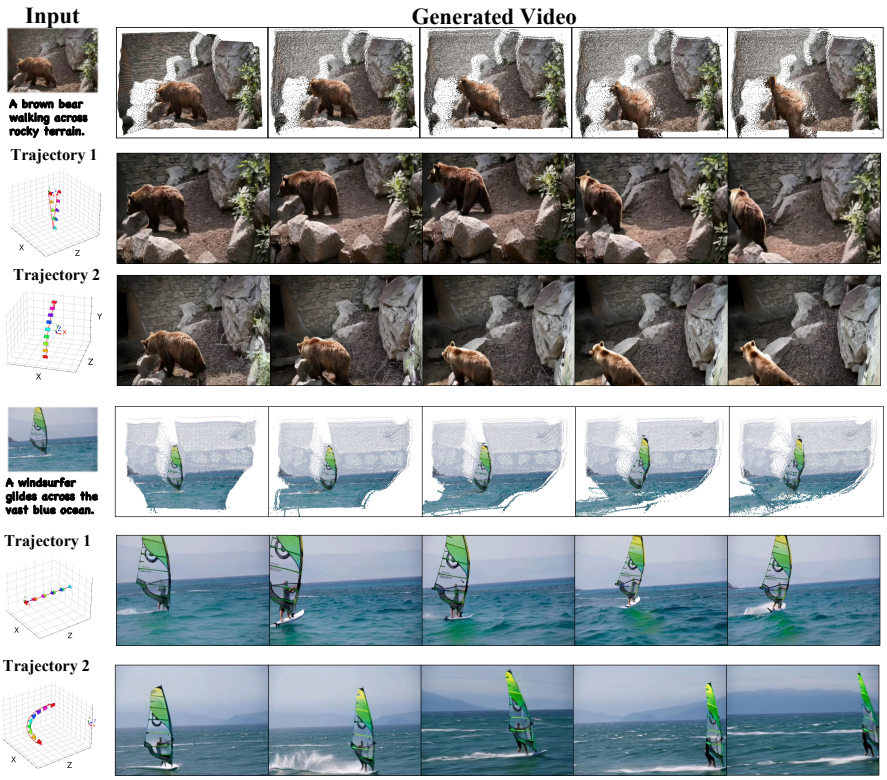

Fig. 4: Qualitative results of our model. The first row shows the 4D point cloud generated by our 4D-STraG. The second and third rows show the videos rendered by our 4D-ViSM under two distinct, user-defined camera trajectories.

5.2 Qualitative Results

For qualitative validation of our model effectiveness, we conduct comparative analyses in Figure 4 and Figure 5. Figure 4 presents two representative scenes, with the first row displaying the point cloud outputs of 4D-STraG. It can be observed that the 4D point clouds maintain structural consistency and exhibit reasonable motion over time, with high detail completeness. For each scene, we define two camera trajectories and perform rendering using 4D-ViSM. The results demonstrate that the rendered images not only preserve the geometric accuracy of the original structure but also effectively align with the camera trajectories, reflecting strong visual consistency. Figure 5 further compares the visual outcomes of several existing methods, including 4Real, DimensionX, Gen3C, and Free4D. In comparison to these approaches, our method generates more diverse and realistic motion, validating the advantage of the our strategy in ensuring both structural rationality and motion plausibility for 4D generation.

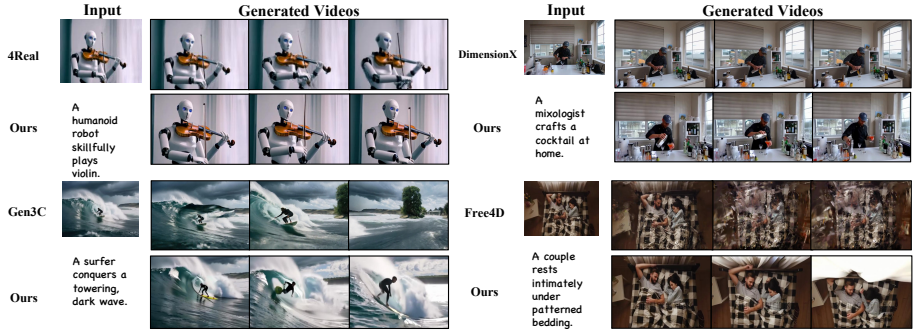

Fig. 5: Qualitative comparison with baseline methods. For each sample, the first row shows the baseline results while the second row presents our MoGe4D results. The first column displays the input image and text prompt.

Table 1: Quantitative comparison on VBench. Higher values are better. The best results in each comparison group are marked in **bold**. Paradigm: Opt. = Optimization-based, Learn. = Learning-based, TF = Training-free.

Exp. No.	Model / Metrics	Paradigm	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
I	4Real [72]	Opt.	**0.9329**	**0.9709**	0.9664	0.7708	0.4938	0.5095
	MoGe4D (Ours)	Learn.	0.8752	0.9364	**0.9682**	**1.0000**	**0.5613**	**0.6230**
II	GenXD [75]	Learn.	0.8042	0.8789	0.9030	**1.0000**	0.4077	0.5209
	DimensionX [51]	Learn.	0.7553	0.8481	**0.9827**	**1.0000**	0.4634	0.5545
	MoGe4D (Ours)	Learn.	**0.8241**	**0.9044**	0.9760	0.9500	**0.4820**	**0.5828**
III	Free4D [37]	TF	0.7899	0.8883	0.9797	**1.0000**	0.3607	0.3562
	Gen3C [48]	Learn.	0.8112	0.8871	**0.9845**	0.9940	0.3812	0.4814
	MoGe4D (Ours)	Learn.	**0.8339**	**0.9065**	0.9773	0.9000	**0.4820**	**0.5939**

5.3 Quantitative Results

Video Quality Benchmarking. Following prior 4D generation works [37], we use VBench [23] for a fair and extensive video-quality comparison against a wider range of baselines. A more detailed 4D geometric analysis is provided in the supplementary material. We compare it with leading 4D generation methods: 4Real, GenXD, DimensionX, Gen3C, and Free4D. Following the evaluation protocol in Free4D [37], we structured our comparisons into three groups based on model availability and technical constraints, with each group evaluated under appropriate camera trajectories. In Group I, we compared against the closed-source 4Real using simple trajectories, as we could only access their official demonstration videos for evaluation. Group II benchmarks single-image 3D models (GenXD, DimensionX's S-Director) with moderate 90° left rotations, due to limited trajectory support in their open-source code. For Group III, we evaluated against other 4D generation models using complex trajectories, including upward, forward, leftward, rightward, and downward 90° movements, to assess performance under challenging camera motions. For Groups II and III, we used 200 held-out samples from WebVid-10M that do not overlap with our TrajScene-60k training subset to ensure fair evaluation in diverse scenarios. As shown in Table 1, MoGe4D consistently achieved promising results across groups. Notably,

Table 2: Quantitative evaluation on 4D Generation Consistency via VLM-based assessment. Higher values are better.

Exp. No.	Model / Metrics	3D Geometric Consistency	Temporal Texture Stability	Subject Identity Preservation	Motion Geometry Coupling	Background Stability Consistency	Average Score
I	4Real [72]	3.04	3.49	4.11	2.64	3.49	3.35
	MoGe4D (Ours)	**3.46**	**3.88**	**4.42**	**3.25**	**3.96**	**3.80**
II	GenXD [75]	2.43	2.66	3.09	2.38	2.65	2.64
	DimensionX [51]	2.58	2.74	3.35	2.53	2.79	2.80
	MoGe4D (Ours)	**3.53**	**3.92**	**4.33**	**3.45**	**3.93**	**3.83**
III	Free4D [37]	1.13	1.30	1.37	1.17	1.17	1.23
	Gen3C [48]	1.99	2.26	2.54	2.01	2.31	2.22
	MoGe4D (Ours)	**3.47**	**3.88**	**4.25**	**3.35**	**3.85**	**3.76**

it demonstrated superior Dynamics, Aesthetic, and Imaging Quality compared to 4Real, outperformed 3D reconstruction baselines in consistency and visual quality, and maintained a pronounced lead in Aesthetic and Imaging Quality against 4D generation models under complex trajectories. Furthermore, consistent improvements across all groups demonstrate our robust generalization.

4D Consistency Analysis. Because standard metrics like VBench insufficiently penalize 3D-specific artifacts (e.g., geometric drift or texture swimming), we introduce a VLM-based protocol for comprehensive 4D evaluation. For videos generated across the aforementioned three groups, we uniformly sample 8 frames and prompt Qwen2.5-VL-72B-Instruct [6] to rate them on a 1-5 scale across five criteria: 3D Geometric Consistency, Temporal Texture Stability, Subject Identity, Motion-Geometry Coupling, and Background Stability. We provide the exact scoring prompt and frame sampling strategy in the supplementary material for full reproducibility. As Table 2 shows, MoGe4D consistently outperforms all baselines. Notably, our significant margins in Motion-Geometry Coupling and Temporal Texture Stability validate MoGe4D's superior spatial-temporal coherence and structural preservation during complex camera motions.

Moreover, to explicitly quantify structural stability, we conduct two rigorous geometry-aware evaluations: (1) **Avg. Trajectory Error**, measuring the L_2 distance between point trajectories tracked by DELTA in generated videos versus their ground-truth counterparts (evaluated on all generated samples from Groups II and III following DELTA's default tracking settings); and (2) **Avg. 3D Reprojection Error**, which follows WorldScore [74] by utilizing DROID-SLAM [53] to gauge the geometric deviation across all co-visible pixels in consecutive frames. As shown in Table 3a, MoGe4D achieves competitive 3D consistency. Notably, our method reduces the trajectory error by a substantial margin while maintaining highly competitive or superior reprojection errors. These results explicitly validate MoGe4D's exceptional pixel-level tracking accuracy and robust rigid 3D spatial coherence.

Inference Efficiency. As Table 3b shows, unlike many optimization-based baselines that require hours of computation, MoGe4D achieves highly competitive inference speed. On a single A100 GPU, it synthesizes 49-frame, 512×368 sequences in just 6 minutes. While GenXD is faster (2 mins), it yields significantly shorter, lower-resolution outputs. Crucially, we are $5\times$ faster than Free4D under

Table 3: Runtime and error comparison. Left: runtime on a single NVIDIA A100 (avg. 100 samples). Right: trajectory & reprojection errors (lower is better)

(a) Mean Trajectory & Reprojection Errors

Exp.	Model	Trajectory Error ↓	Reprojection Error ↓
i	Ground truth	0.000	0.301
	GenXD [75]	0.236	1.891
	DimensionX [51]	0.465	**0.602**
	MoGe4D (Ours)	**0.058**	0.614
ii	Ground truth	0.000	0.284
	Free4D [37]	0.252	0.652
	Gen3C [48]	0.197	0.755
	MoGe4D (Ours)	**0.042**	**0.639**

(b) Runtime Analysis(on a single A100)

Method	Year	Resolution	Frames	Time
4Dfy [4]	CVPR'24	256×256	–	10 h
4Real [72]	NeurIPS'24	256×144	8	1.5 h
Gen3C [48]	CVPR'25	1280×704	121	50 min
GenXD [75]	ICLR'25	256×256	12	2 min
Free4D [37]	ICCV'25	512×368	16	30 min
MoGe4D (Ours)	–	512×368	49	6 min

Table 4: Component-wise Ablation Study. We analyze the impact of depth normalization, latent design, MPM structure, and training data variations. Best results per metric are highlighted in **bold**.

Metric	w/o Depth Norm	w/o Depth Latents	MPM Module		Training Data			MoGe4D
			w/o MPM	w/o Patch Feat.	Small Scale	Low Quality	Noise	
Consistency↑	0.8604	0.8567	0.8650	**0.8743**	0.8567	0.8549	0.8685	0.8702
Dynamic↑	0.8850	0.8500	0.8500	0.8840	0.8920	0.8850	0.8967	**0.9000**
Aesthetic↑	0.4672	0.4738	0.4806	0.4754	0.4791	0.4654	0.4771	**0.4820**

equivalent resolutions. This remarkable efficiency allows our pipeline to inherently support longer-duration video synthesis.

5.4 Ablation Studies

Table 4 presents a comprehensive ablation study on VBench under the Group III setting. Removing depth-guided normalization drops Consistency to 0.8604 and Aesthetic quality to 0.4672, while excluding depth latents degrades both Consistency and Dynamic scores, as the model struggles to maintain motion coherence. Removing MPM reduces Dynamic from 0.9 to 0.85; replacing its patch-level features with a global `[CLS]` token slightly improves Consistency but degrades Dynamic and Aesthetic quality, confirming that fine-grained patch-level conditioning is essential for high-fidelity 4D generation. For training data, reducing scale to a 1k subset or using unfiltered 60k samples both cause significant performance drops, validating that large-scale, high-quality data is foundational. Notably, adding random noise during training yields only negligible degradation, demonstrating robustness to noisy pseudo-GT supervision. Our full MoGe4D achieves optimal performance across all metrics, validating the synergistic contribution of each component.

Row 1-6 in Figure 6 visually demonstrate these findings. Baseline min-max normalization (rows 1-2) yields unstable point clouds with exaggerated movements, particularly in scenes with large depth variations. Our depth-guided approach effectively resolves this issue. Further visualizations (rows 3-6) show that without MPM or depth guidance, the model fails to preserve object structure or maintain motion consistency across different parts, confirming the critical role of depth in ensuring structurally coherent point trajectories.

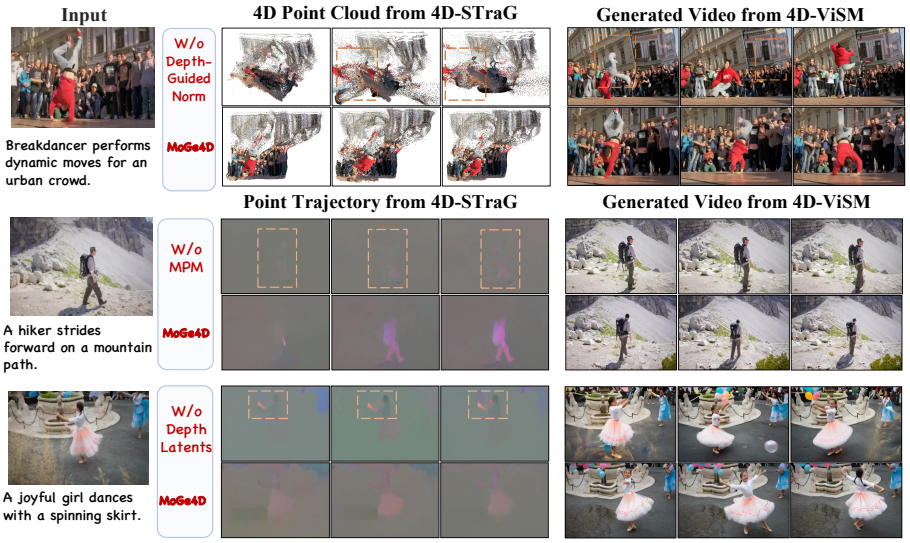

Fig. 6: Ablation studies on normalization methods and module components. (Rows 1-2) Depth-guided motion normalization stabilizes 4D point cloud generation. (Rows 3-6) Removing the MPM module reduces motion magnitude while excluding depth guidance breaks structural motion consistency, validating our design choices.

Fig. 7: Limitations: Extreme lighting and unstructured crowds.

5.5 Discussions

Geometry-Conditioned vs. Sequential 4D Generation. To validate our geometry-conditioned trajectory generation paradigm, we compare against two sequential baselines: Wan2.1-I2V [57] followed by either DELTA [44] tracking or VGGT [58] reconstruction. As shown in Figure 8, sequential pipelines suffer from severe error accumulation: since video diffusion models lack explicit 3D constraints, frame-level appearance inconsistencies are misinterpreted as 3D motion by downstream modules, producing spurious background drift and geometric fragmentation. In contrast, our approach conditions trajectory generation directly on the initial point cloud geometry, enabling the diffusion model to produce temporally coherent, geometry-consistent 4D trajectories without relying on error-prone intermediate video synthesis.

Limitations. As shown in Fig. 7, MoGe4D struggles with extreme photometric conditions (e.g., harsh backlighting) and highly unstructured dynamics (e.g., chaotic crowds). In these complex edge cases, off-the-shelf monocular depth estimators often degrade, introducing noise into the initial geometric prior. Since our dense trajectory generation is conditionally bounded by the quality of this initial

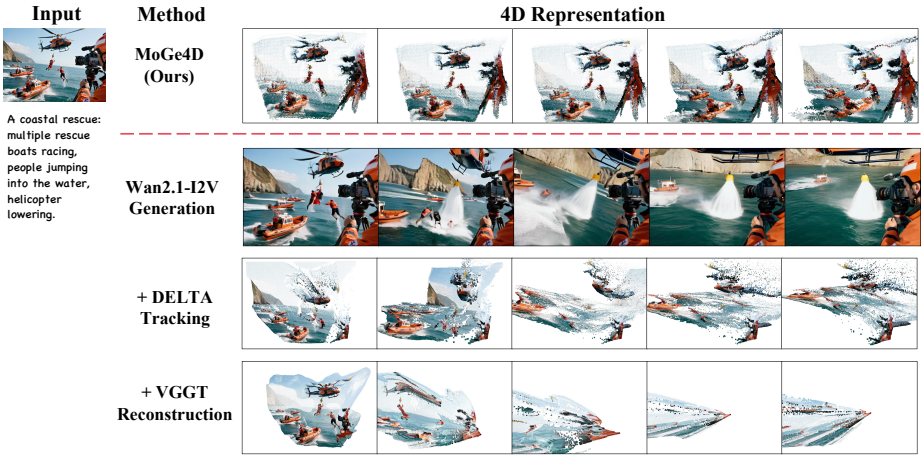

Fig. 8: Geometry-conditioned 4D-STraG (top row) vs. sequential pipelines. Rows 2–4 show cascaded baselines: Wan2.1-I2V video generation followed by DELTA tracking or VGGT reconstruction. By conditioning trajectory generation directly on initial point cloud geometry, our method achieves superior spatio-temporal coherence, whereas sequential approaches suffer fragmentation and drift from error accumulation. All results are rendered from a fixed camera viewpoint for fair comparison.

geometry, the resulting motion can become less reliable. Integrating more robust visual foundation models to handle such extreme physical conditions remains an important direction for future work.

6 Conclusion

We proposed MoGe4D, a geometry-conditioned framework for single-image 4D synthesis that represents a scene as dense time-varying point trajectories, overcoming weaknesses of decoupled paradigms. To support training, we constructed TrajScene-60K, a large-scale dataset with dense 4D trajectories, and introduced the 4D-STraG diffusion model with depth-guided motion normalization and motion perception priors. Experiments show that MoGe4D achieves superior geometric consistency, dynamic realism, and visual fidelity compared to existing approaches. This work establishes a new technical pathway for 4D synthesis from minimal input, and points toward future exploration of more universal dynamic priors and lightweight 4D representations for practical deployment.

7 Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant 62336004, Grant 623B2063, Grant 62125603, Grant 62576188, and Grant 62321005.

References

1. Alibaba PAI Team: Wan2.1-Fun-14B-Control. https://huggingface.co/alibaba-pai/Wan2.1-Fun-14B-Control (2024), https://huggingface.co/alibaba-pai/Wan2.1-Fun-14B-Control 9
2. Alibaba PAI Team: Wan2.1-Fun-14B-InP. https://huggingface.co/alibaba-pai/Wan2.1-Fun-14B-InP (2024), https://huggingface.co/alibaba-pai/Wan2.1-Fun-14B-InP 9
3. Babaeizadeh, M., Saffar, M.T., Nair, S., Levine, S., Finn, C., Erhan, D.: Fitvid: Overfitting in pixel-level video prediction. arXiv preprint arXiv:2106.13195 (2021) 3
4. Bahmani, S., Skorokhodov, I., Rong, V., Wetzstein, G., Guibas, L., Wonka, P., Tulyakov, S., Park, J.J., Tagliasacchi, A., Lindell, D.B.: 4D-fy: Text-to-4d generation using hybrid score distillation sampling. In: CVPR. pp. 7996–8006 (2024) 3, 4, 13
5. Bai, J., Xia, M., Fu, X., Wang, X., Mu, L., Cao, J., Liu, Z., Hu, H., Bai, X., Wan, P., et al.: Recammaster: Camera-controlled generative rendering from a single video. arXiv preprint arXiv:2503.11647 (2025) 4
6. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 12
7. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV. pp. 1728–1738 (2021) 4
8. Bian, W., Huang, Z., Shi, X., Li, Y., Wang, F.Y., Li, H.: Gs-dit: Advancing video generation with pseudo 4d gaussian fields through efficient dense 3d point tracking. arXiv preprint arXiv:2501.02690 (2025) 2, 4
9. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR. pp. 19457–19467 (2024) 4
10. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation. arXiv preprint arXiv:2310.19512 (2023) 3
11. Chen, Z., Liu, T., Zhuo, L., Ren, J., Tao, Z., Zhu, H., Hong, F., Pan, L., Liu, Z.: 4dnex: Feed-forward 4d generative modeling made easy. arXiv preprint arXiv:2508.13154 (2025) 4
12. Chen, Z., Cao, J., Chen, Z., Li, Y., Ma, C.: Echomimic: Lifelike audio-driven portrait animations through editable landmark conditions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 2403–2410 (2025) 3
13. Clark, A., Donahue, J., Simonyan, K.: Adversarial video generation on complex datasets. arXiv preprint arXiv:1907.06571 (2019) 3
14. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. arXiv preprint arXiv:2412.14169 (2024) 3
15. Denton, E., Fergus, R.: Stochastic video generation with a learned prior. In: ICML. pp. 1174–1183. PMLR (2018) 3
16. Feng, H., Zhang, J., Wang, Q., Ye, Y., Yu, P., Black, M.J., Darrell, T., Kanazawa, A.: St4rtrack: Simultaneous 4d reconstruction and tracking in the world. In: ICCV. pp. 8503–8513 (2025) 4
17. Gao, R., Holynski, A., Henzler, P., Brussee, A., Martin-Brualla, R., Srinivasan, P., Barron, J.T., Poole, B.: Cat3d: Create anything in 3d with multi-view diffusion models. arXiv preprint arXiv:2405.10314 (2024) 4

18. Girdhar, R., El-Nouby, A., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Omnimae: Single model masked pretraining on images and videos. In: CVPR. pp. 10406–10417 (2023) 8
19. Guo, C., Jiang, T., Chen, X., Song, J., Hilliges, O.: Vid2avatar: 3d avatar reconstruction from videos in the wild via self-supervised scene decomposition. In: CVPR. pp. 12858–12868 (2023) 2
20. He, J., Lehrmann, A., Marino, J., Mori, G., Sigal, L.: Probabilistic video generation using holistic attribute control. In: ECCV. pp. 452–467 (2018) 3
21. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. NeurIPS **35**, 8633–8646 (2022) 3
22. Hong, W., Wang, W., Ding, M., Yu, W., Lv, Q., Wang, Y., Cheng, Y., Huang, S., Ji, J., Xue, Z., et al.: Cogvlm2: Visual language models for image and video understanding. arXiv preprint arXiv:2408.16500 (2024) 4
23. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., et al.: Vbench: Comprehensive benchmark suite for video generative models. In: CVPR. pp. 21807–21818 (2024) 11
24. Jiang, Y., Yu, C., Cao, C., Wang, F., Hu, W., Gao, J.: Animate3d: Animating any 3d model with multi-view video diffusion. NeurIPS **37**, 125879–125906 (2024) 2, 4
25. Jiang, Y., Zhang, L., Gao, J., Hu, W., Yao, Y.: Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. arXiv preprint arXiv:2311.02848 (2023) 4
26. Jiang, Z., Zheng, C., Laina, I., Larlus, D., Vedaldi, A.: Geo4d: Leveraging video generators for geometric 4d scene reconstruction. arXiv preprint arXiv:2504.07961 (2025) 4
27. Jin, I.H., Choo, H., Jeong, S.H., Heemoon, P., Kim, J., Kwon, O.j., Kong, K.: Optimizing 4d gaussians for dynamic scene video from single landscape images. In: ICLR (2025) 2
28. Karaev, N., Makarov, Y., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: ICCV. pp. 6013–6022 (2025) 4
29. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: ECCV. pp. 18–35. Springer (2024) 4
30. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023) 5
31. Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., Shi, H.: Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In: ICCV. pp. 15954–15964 (2023) 3
32. Lin, C., Lin, Y., Pan, P., Yu, Y., Yan, H., Fragkiadaki, K., Mu, Y.: Movies: Motion-aware 4d dynamic view synthesis in one second. arXiv preprint arXiv:2507.10065 (2025) 4
33. Lin, C.H., Lv, Z., Wu, S., Xu, Z., Nguyen-Phuoc, T., Tseng, H.Y., Straub, J., Khan, N., Xiao, L., Yang, M.H., et al.: Dgs-lrm: Real-time deformable 3d gaussian reconstruction from monocular videos. arXiv:2506.09997 (2025) 4
34. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023) 8
35. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024) 4
36. Liu, F., Sun, W., Wang, H., Wang, Y., Sun, H., Ye, J., Zhang, J., Duan, Y.: Reconx: Reconstruct any scene from sparse views with video diffusion model. arXiv preprint arXiv:2408.16767 (2024) 4

37. Liu, T., Huang, Z., Chen, Z., Wang, G., Hu, S., Shen, L., Sun, H., Cao, Z., Li, W., Liu, Z.: Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. arXiv preprint arXiv:2503.20785 (2025) 2, 4, 11, 12, 13
38. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR 9
39. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: AAAI. vol. 38, pp. 4117–4125 (2024) 3
40. Mi, Z., Wang, Y., Xu, D.: One4d: Unified 4d generation and reconstruction via decoupled lora control. arXiv preprint arXiv:2511.18922 (2025) 4
41. Miao, Q., Li, K., Quan, J., Min, Z., Ma, S., Xu, Y., Yang, Y., Liu, P., Luo, Y.: Advances in 4d generation: A survey. arXiv preprint arXiv:2503.14501 (2025) 2
42. Miao, Q., Quan, J., Li, K., Luo, Y.: Pla4d: Pixel-level alignments for text-to-4d gaussian splatting. arXiv preprint arXiv:2405.19957 (2024) 2
43. Müller, N., Schwarz, K., Rössle, B., Porzi, L., Bulo, S.R., Nießner, M., Kontschieder, P.: Multidiff: Consistent novel view synthesis from a single image. In: CVPR. pp. 10258–10268 (2024) 4
44. Ngo, T.D., Zhuang, P., Kalogerakis, E., Gan, C., Tulyakov, S., Lee, H.Y., Wang, C.: Delta: Dense efficient long-range 3d tracking for any video. In: ICLR (2025) 4, 5, 14
45. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023) 8
46. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. arXiv preprint arXiv:2502.20110 (2025) 7, 9
47. Ren, J., Xie, C., Mirzaei, A., Kreis, K., Liu, Z., Torralba, A., Fidler, S., Kim, S.W., Ling, H., et al.: L4gm: Large 4d gaussian reconstruction model. Advances in Neural Information Processing Systems **37**, 56828–56858 (2025) 4
48. Ren, X., Shen, T., Huang, J., Ling, H., Lu, Y., Nimier-David, M., Müller, T., Keller, A., Fidler, S., Gao, J.: Gen3c: 3d-informed world-consistent video generation with precise camera control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6121–6132 (2025) 2, 4, 11, 12, 13
49. Shi, X., Huang, Z., Wang, F.Y., Bian, W., Li, D., Zhang, Y., Zhang, M., Cheung, K.C., See, S., Qin, H., et al.: Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In: ACM SIGGRAPH. pp. 1–11 (2024) 3
50. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: ICLR (2023) 3
51. Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., Wang, Y.: Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928 (2024) 2, 4, 11, 12, 13
52. Team, K.A.: Kling image-to-video model (2024), https://klingai.com/image-to-video/ 3
53. Teed, Z., Deng, J.: Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. NeurIPS **34**, 16558–16569 (2021) 12
54. Tulyakov, S., Liu, M.Y., Yang, X., Kautz, J.: Mocogan: Decomposing motion and content for video generation. In: CVPR. pp. 1526–1535 (2018) 3
55. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. NeurIPS **30** (2017) 3

56. Vondrick, C., Pirsiavash, H., Torralba, A.: Generating videos with scene dynamics. vol. 29 (2016) 3
57. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025) 3, 7, 9, 14
58. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5294–5306 (2025) 14
59. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: ICCV. pp. 9660–9672 (2025) 4
60. Wu, D., Liu, F., Hung, Y.H., Qian, Y., Zhan, X., Duan, Y.: 4d-fly: Fast 4d reconstruction from a single monocular video. In: CVPR. pp. 16663–16673 (2025) 4
61. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: CVPR. pp. 20310–20320 (2024) 4
62. Wu, J., Li, X., Zeng, Y., Zhang, J., Zhou, Q., Li, Y., Tong, Y., Chen, K.: Motionbooth: Motion-aware customized text-to-video generation. NeurIPS **37**, 34322–34348 (2024) 3
63. Wu, R., Gao, R., Poole, B., Trevithick, A., Zheng, C., Barron, J.T., Holynski, A.: Cat4d: Create anything in 4d with multi-view video diffusion models. In: CVPR. pp. 26057–26068 (2025) 2, 4
64. Wu, Z., Yu, C., Jiang, Y., Cao, C., Wang, F., Bai, X.: Sc4d: Sparse-controlled video-to-4d generation and motion transfer. In: ECCV. pp. 361–379. Springer (2024) 4
65. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. arXiv:2507.12462 (2025) 4
66. Xiao, Y., Wang, Q., Zhang, S., Xue, N., Peng, S., Shen, Y., Zhou, X.: Spatialtracker: Tracking any 2d pixels in 3d space. In: CVPR. pp. 20406–20417 (2024) 4
67. Xie, Y., Yao, C.H., Voleti, V., Jiang, H., Jampani, V.: Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. arXiv preprint arXiv:2407.17470 (2024) 4
68. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. IEEE TVCG **31**(2), 1526–1541 (2024) 3
69. Yang, S., Hou, L., Huang, H., Ma, C., Wan, P., Zhang, D., Chen, X., Liao, J.: Direct-a-video: Customized video generation with user-directed camera movement and object motion. In: ACM SIGGRAPH 2024 Conference Papers. pp. 1–12 (2024) 3
70. Yao, C.H., Xie, Y., Voleti, V., Jiang, H., Jampani, V.: Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. In: ICCV. pp. 13248–13258 (2025) 4
71. Yao, D.Y., Zhai, A.J., Wang, S.: Uni4d: Unifying visual foundation models for 4d modeling from a single video. In: CVPR. pp. 1116–1126 (2025) 4
72. Yu, H., Wang, C., Zhuang, P., Menapace, W., Siarohin, A., Cao, J., Jeni, L., Tulyakov, S., Lee, H.Y.: 4Real: Towards photorealistic 4d scene generation via video diffusion models. NeurIPS **37**, 45256–45280 (2024) 2, 4, 11, 12, 13
73. Yu, W., Xing, J., Yuan, L., Hu, W., Li, X., Huang, Z., Gao, X., Wong, T.T., Shan, Y., Tian, Y.: Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048 (2024) 4

74. Zhang, Q., Zhai, S., Martin, M.A.B., Miao, K., Toshev, A., Susskind, J., Gu, J.: World-consistent video diffusion with explicit 3d modeling. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 21685–21695 (2025) 12
75. Zhao, Y., Lin, C.C., Lin, K., Yan, Z., Li, L., Yang, Z., Wang, J., Lee, G.H., Wang, L.: GenXD: Generating Any 3D and 4D Scenes. In: ICLR (2025) 11, 12, 13
76. Zhao, Y., Yan, Z., Xie, E., Hong, L., Li, Z., Lee, G.H.: Animate124: Animating one image to 4d dynamic scene. arXiv preprint arXiv:2311.14603 (2023) 2, 4
77. Zhou, K., Li, W., Wang, Y., Hu, T., Jiang, N., Han, X., Lu, J.: Nerflix: High-quality neural view synthesis by learning a degradation-driven inter-viewpoint mixer. In: CVPR. pp. 12363–12374 (2023) 4
78. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. In: ECCV. pp. 145–163. Springer (2024) 4