arXiv:2510.26796v1 [cs.CV] 30 Oct 2025

See4D: Pose-Free 4D Generation via Auto-Regressive Video Inpainting

DONGYUE LU* and AO LIANG*, NUS, Singapore
TIANXIN HUANG, HKU, Hong Kong SAR
XIAO FU, CUHK, Hong Kong SAR
YUYANG ZHAO, NUS, Singapore
BAORUI MA, Tsinghua, China
LIANG PAN, Shanghai AI Laboratory, China
WEI YIN, Horizon Robotics, China
LINGDONG KONG†, NUS, Singapore
WEI TSANG OOI, NUS, Singapore
ZIWEI LIU, NTU, Singapore

Fig. 1. **Illustration of our 4D generation framework** (See4D). Given an unposed source video, we spline-interpolate a trajectory of virtual camera poses and estimate per-frame depth to lift each frame into 3D. Depth-guided forward warping produces intermediate latents which, together with source latents, are processed by our view-conditional diffusion model, augmented with spatiotemporal attention and noise-adaptive conditioning in an auto-regressive framework, synthesizing the final target-view sequence. For demo examples, see **Webpage** at https://see-4d.github.io.

Immersive applications call for synthesizing spatiotemporal 4D content from casual videos without costly 3D supervision. Existing video-to-4D methods typically rely on manually annotated camera poses, which are labor-intensive and brittle for in-the-wild footage. Recent warp-then-inpaint approaches mitigate the need for pose labels by warping input frames along a novel camera trajectory and using an inpainting model to fill missing regions, thereby depicting the 4D scene from diverse viewpoints. However, this trajectory-to-trajectory formulation often entangles camera motion with scene dynamics and complicates both modeling and inference. We introduce See4D, a pose-free, trajectory-to-camera framework that replaces explicit trajectory prediction with rendering to a bank of fixed virtual cameras, thereby separating camera control from scene modeling. A view-conditional video inpainting model is trained to learn a robust geometry prior by denoising realistically synthesized warped images and to inpaint occluded or missing regions across virtual viewpoints, eliminating the need for explicit 3D annotations. Building on this inpainting core, we design a spatiotemporal autoregressive inference pipeline that traverses virtual-camera splines and extends videos with overlapping windows, enabling coherent generation at

*Both authors contributed equally to this research.
†Project Lead.

Authors' Contact Information: Dongyue Lu, dongyue.lu@u.nus.edu; Ao Liang, a_liang@u.nus.edu, NUS, Singapore; Tianxin Huang, HKU, Hong Kong SAR, 21725129@zju.edu.cn; Xiao Fu, CUHK, Hong Kong SAR, fx023@ie.cuhk.edu.hk; Yuyang Zhao, NUS, Singapore, yuyangzhao98@gmail.com; Baorui Ma, Tsinghua, China, mabaorui2014@gmail.com; Liang Pan, Shanghai AI Laboratory, China, pan.liang@u.nus.edu; Wei Yin, Horizon Robotics, China, yvanwy@outlook.com; Lingdong Kong, NUS, Singapore, lingdong.kong@u.nus.edu; Wei Tsang Ooi, NUS, Singapore, ooiwt@comp.nus.edu.sg; Ziwei Liu, NTU, Singapore, ziwei.liu@ntu.edu.sg.

bounded per-step complexity. We validate See4D on cross-view video generation and sparse reconstruction benchmarks. Across quantitative metrics and qualitative assessments, our method achieves superior generalization and improved performance relative to pose- or trajectory-conditioned baselines, advancing practical 4D world modeling from casual videos.

1 Introduction

Virtual Reality (VR) has evolved from a niche curiosity into a pervasive medium that impacts storytelling, scientific communication, and professional training [10, 77, 100]. A truly immersive experience, however, demands more than static panoramic images. It requires the ability to wander freely through a scene as it unfolds over time, i.e., a 4D environment that is simultaneously spatially and temporally consistent [1]. Delivering such content calls for densely synchronized, multi-view videos captured with tightly calibrated camera arrays [40]. These systems are costly, fragile, and generally impractical outside controlled studios, forcing most consumer VR experiences to rely on sparse or synthetic assets. Bridging this gap by converting a single, casually captured video into a coherent 4D representation would unlock an enormous amount of real-world footage for VR content creation, motivating the recent surge of research in **video-to-4D** generation.

Current efforts address this problem along two principal directions. The first class of methods conditions the model on **explicit poses**, assuming that each frame is paired with a precise external pose, and trains the network to generate novel views under these specified poses [2, 3, 79, 93, 99]. While effective in simulation, these methods rely on explicit pose annotations, which are difficult to obtain in the dynamic, hand-held videos common on online platforms. The second class leverages monocular depth estimation and employs a **warp-then-inpaint** strategy to synthesize novel views [22, 32, 52, 63, 92]. Pixels are first projected into the target view using estimated depth [30, 83], yielding an incomplete approximation of the target view constrained by visible geometry. An inpainting model, conditioned on the **warped image**, is then employed to hallucinate the missing content. Compared to pose-based conditioning, warped-image conditioning generally achieves stronger 3D consistency by explicitly encoding geometric cues and preserving cross-view coherence.

Despite this superiority, most warp-then-inpaint systems are optimized for **trajectory-to-trajectory** synthesis: given a video captured along a specific camera path, they synthesize a novel video from an entirely different trajectory, preserving synchronized scene dynamics [2, 3, 32, 63, 69, 92]. This dual-trajectory formulation, however, is relatively under-constrained, which can complicate training and lead to less stable inference. The generated viewpoints often remain close to the input path, and the reconstruction quality can be sensitive to per-frame depth accuracy. Consequently, these approaches face challenges in scaling to full 4D scene reconstruction that enables flexible free-viewpoint playback, a capability that is particularly important for immersive VR experiences.

To address this challenge, we propose See4D, a pose-free framework based on a **trajectory-to-camera** formulation. Our method generates a bank of synchronized, fixed-view videos from a single monocular clip. We train a view-conditional inpainting model on depth-warped images and associated masks instead of relying on explicit camera poses, and integrate a lightweight spatial-temporal transformer backbone for cross-view and cross-frame coherence. During training, we apply *realistic warp synthesis* by forward-projecting the target-view video to a nearby random pose and back-projecting with jitter, simulating the source-to-target warp artifacts observed during inference. Additionally, we propose *noise-adaptive condition*, which adds modulated noise to the conditional warped images based on their corresponding warp mask, preventing overfitting to unreliable warps. This design simplifies optimization compared to traditional **trajectory-to-trajectory** methods, remains robust in dynamic and complex scenarios, and produces high-quality, geometrically consistent novel views suitable for large-scale 4D scene modeling and representation.

Complementing the core model, we introduce a **spatio-temporal auto-regressive** inference pipeline that robustly scales synthesis to long sequences and viewpoint shifts. Spatially, we march smoothly along a spline of virtual camera poses between the source and each target view, iteratively warping, inpainting, and updating depth to progressively mitigate cross-view self-occlusions. Temporally, we extend clips by overlapping successive diffusion runs, reusing a sliding window of one prediction as the seed for the next. This dual recursion effectively enforces smooth motion continuity while gracefully handling depth noise and complex disocclusion.

We evaluate our method on both 4D scene reconstruction accuracy and perceptual video quality. On public 4D reconstruction benchmarks, our method delivers the highest completeness and geometric fidelity, surpassing all previous methods. On our newly collected set of challenging hand-held videos, it also wins across standard generation metrics, achieving sharper detail, stable cross-view consistency, and stronger temporal coherence than prior work. Beyond numbers, we demonstrate practical benefits in robot manipulation, autonomous driving, interactive gaming, and movie production, underlining the real-world value of our pipeline.

The key contributions of this work can be summarized as:

- We propose See4D, a pose-free pipeline that learn to generate 4D scenes directly from in-the-wild videos, eliminating the need for large-scale pose annotations.
- We propose a trajectory-to-camera formulation. By predicting a set of fixed target views rather than a full camera path, we simplify training and improve synthesis stability.
- We introduce a spatial-temporal auto-regressive inference process that delivers long, temporally consistent videos while coping with self-occlusion and depth noise.
- Extensive evaluations show that our method outperforms existing baselines in both reconstruction accuracy and generative quality, unlocking new workflows for VR content creation.

2 Related Work

2.1 3D & 4D Reconstruction

Early neural rendering methods exploit either sparse multi-view supervision or single-image priors to reconstruct static scenes, using NeRF-style implicit fields [5–7, 24, 34, 39, 55, 65, 76] and, more recently, Gaussian-splatting variants [11, 23, 31, 37, 51, 82, 94, 101].

To mitigate the ill-posedness from limited viewpoints, follow-up work fuses classical geometry with learned priors, either by bootstrapping depth, optical flow, or patch correspondences [14, 18, 84, 95] or by regularizing the radiance field with geometric constraints [42, 56, 62, 72, 84]. Parallel efforts replace slow per-scene optimization with feed-forward prediction of point clouds, meshes, or Gaussians [68, 75], trading some fidelity for speed. Extending these ideas to the temporal dimension yields 4D reconstruction. NeRF-style methods [15, 45, 46, 59, 60, 62, 70, 81, 90] and dynamic Gaussian splats [16, 19, 78, 86, 88] faithfully capture moderate motion when multi-view videos are available. However, monocular settings remain considerably harder; for example, Shape-of-Motion [73] combines depth, tracking, and mask priors, yet still leaves holes in occluded regions. Recent systems therefore lift monocular frames into pseudo-multi-view observations that a downstream reconstructor can absorb [32, 99]. Despite impressive progress, balancing temporal coherence, geometric accuracy, and capture convenience remains a fundamental open challenge.

2.2 Generative Novel View Synthesis

Diffusion models now act as versatile priors that can be blended with geometry in several ways. Score-distillation pipelines fit a 3D representation such that its renderings match image-level diffusion gradients [13, 61, 102]. Another thread explicitly injects scene structure, including depth maps, meshes, or multi-view images into diffusion backbones, yielding high spatial fidelity but relying on accurate poses and largely static content [20, 37, 47, 48, 52, 66, 75, 76, 80, 91, 93, 99]. Extending these ideas to 4D novel-view synthesis requires modeling object motion and scene-level changes over time. Approaches based on pose-aware diffusion or small-scale dynamics [49, 57, 79, 99] often falter under large motion or complex layouts, as shown by WideRange4D [85]. Complementary efforts refine object trajectories [36] or distill geometry from pretrained 2D/3D models in a tuning-free manner [50], yet broad spatial coverage and strong temporal coherence remain elusive. Parallel progress in video inpainting demonstrates the temporal consistency of diffusion models on inpainting incomplete clips [12, 17, 21, 41, 43, 71, 97, 103], inspiring pipelines that replace costly geometry with warp-then-inpaint recipes. See3D [52] learns an implicit geometry prior from pose-free videos, and BlobCtrl [44] bridges 2D generation with 3D editing via Gaussian blobs. To handle wider baselines, several works [8, 98] inpaint masked views produced by stereo cues or Gaussian fields. Similar strategies combine geometric masks from depth warping or synthetic multi-view data with fine-tuned video diffusion models [9, 28, 58, 63, 87, 89, 92]. While these methods expand view diversity, they rely on costly preprocessing and degrade on long sequences. Achieving wide spatial coverage and stable temporal dynamics without curated multi-view data, therefore, remains open, a gap our pose-free, auto-regressive framework seeks to close.

2.3 Camera-Controlled Video Generation

Recent advances in video diffusion have sparked interest in camera-controlled generation, where users prescribe a viewpoint path and the model renders the scene accordingly [2, 3, 25, 26, 29, 35, 53, 69]. Early solutions often embed 6DoF poses into 2D or video diffusion backbones [2, 63, 99], while TrajectoryCrafter [92] fine-tunes video models on curated multi-view warped data, ensuring the generative network respects the target geometry. CameraCtrl [25] introduces camera classifier-free guidance to handle diverse user trajectories without sacrificing realism, though it often struggles when the camera moves too far from the original viewpoint. Some methods try to address broader exploration ranges. CameraCtrl II [26] combines new training data with a carefully designed pose-injection module to preserve dynamic content over extended camera paths, enabling users to stitch multiple short clips while retaining scene consistency. Meanwhile, Reangle-A-Video [35] implements a video-to-video approach by fine-tuning on warped frames from a single input video, so it can produce cross-view sequences or dynamic camera transitions. GS-DiT [8] and StereoCrafter [98] fuse warped views with diffusion priors, but rely on synthetic training data and remain limited to moderate camera excursions. Such methods mark an important step toward fully interactive 4D generation, letting creators define novel camera trajectories and watch an otherwise static or partially observed scene come to life from all angles.

3 Methodology

We introduce See4D, a pose-free trajectory-to-camera framework that synthesizes consistent 4D scenes from a single monocular video. Unlike trajectory-to-trajectory methods, it renders to a set of fixed virtual cameras, thereby decoupling scene modeling from camera motion and improving stability. Section 3.1 reviews the foundations of latent diffusion and image warping. Section 3.2 presents our view-conditional inpainting model. Section 3.3 describes our auto-regressive inference pipeline that maintains spatial-temporal consistency.

3.1 Preliminaries

3.1.1 Latent Diffusion Model. We adopt a latent diffusion model (LDM) [64] as the generative backbone for our fixed-view synthesis. We first compress an input video x into a latent tensor $z_0 = \mathcal{E}(x)$ using a pre-trained VAE [38] encoder $\mathcal{E}$. A variance-preserving forward process adds Gaussian noise, *i.e.*,

$$z_\tau = \sqrt{\bar{\alpha}_\tau}\, z_0 + \sqrt{1-\bar{\alpha}_\tau}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{1}$$

where $\bar{\alpha}_\tau$ is a fixed scheduler and $\tau \in [1, T]$. A denoiser ε_θ is trained to predict ε by minimizing the following term:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,\varepsilon,\tau}\big[\, \|\varepsilon_\theta(z_\tau, c, \tau) - \varepsilon\|_2^2 \big], \tag{2}$$

where c denotes the controllable condition. At test time, iterative denoiser transforms $z_T \sim \mathcal{N}(0, I)$ back to z_0, which the decoder maps to the output video. Working in latent space cuts computation by an order of magnitude while preserving perceptual quality.

3.1.2 Depth-Guided Image Warp. We warp source frames using predicted depth to build a view-specific prior without explicit poses. Given a per–pixel depth map $\mathbf{D}$, camera intrinsics $\mathbf{K}$, and a relative transform $\mathbf{T} \in \mathrm{SE}(3)$, each pixel $\mathbf{u} = (u, v)$ is lifted to 3D via homogeneous back-projection, mapped to the target frame, and then re-projected. The process is formulated as:

$$\mathbf{u}_w = \pi\big(\mathbf{K}\mathbf{T}\big[\mathbf{D}(\mathbf{u})\mathbf{K}^{-1}[u, v, 1]^T; 1\big]\big), \tag{3}$$

Fig. 2. **Overview of our view-conditional inpainting model**. From a synchronized multi-view dynamic scenario, we select a source–target view pair. We apply realistic warp synthesis (*cf.* Sec.3.2.2) to generate a warp image and mask as conditioning. The warp latent is then perturbed in a noise-adaptive manner (*cf.* Sec.3.2.3). Finally, the source latent, warp condition, and noisy target latent are input to our spatial-transformer backbone (*cf.* Sec.3.2.4) to denoise the latent and inpaint the masked regions. These components yield a pose-free inpainting model stable on dynamic and complex footage.

where $\pi([u, v, w]^T) = [u/w,\ v/w]^T$. We bilinearly splat the source pixels to sub-pixel targets, producing a warped frame $\mathbf{I}_w$ and a mask $\mathbf{M}$ that flags depth clashes or off-screen projections. This sparse yet structured pair conditions the inpainting diffuser, enforcing target-view geometry while letting it hallucinate missing regions.

3.2 View-Conditional Inpainting Model

3.2.1 Overview. Obtaining frame-accurate 6-DoF poses for in-the-wild videos is costly and unreliable. Instead, we adopt the warp-then-inpaint paradigm [52, 63, 92]. During training, given paired source and target videos $(\mathbf{I}_s, \mathbf{I}_t)$, for each target frame $\mathbf{I}_t$, we estimate its depth and warp it to a virtual pose using Eq. (3). This produces a pixel-aligned warped image $\mathbf{I}_w$ together with a visibility mask $\mathbf{M}$. We then construct the conditioning tensor c_τ by fusing $\mathbf{I}_w$, $\mathbf{M}$, and the step-dependent noisy latent z_τ. The denoiser is trained with the objective

$$\mathcal{L} = \mathbb{E}_{z_0, z_s, \varepsilon, \tau}\left[\, \|\varepsilon_\theta(z_\tau, z_s, c_\tau, \tau) - \varepsilon\|_2^2 \right], \tag{4}$$

where z_s is the clean latent of the source video.

Prior warp-conditioned models [32, 52, 92] are tuned for static scenes but fail under fast motion or viewpoint variations. We address this with: **(1) Realistic Warp Synthesis.** To simulate real-world warp artifacts observed during source-to-target projection at inference time, we perturb the target view by forward-projecting it to a randomly sampled nearby virtual camera pose, followed by back-projection using the jittered pose during training. **(2) Noise-Adaptive Condition.** We introduce adaptive noise to the conditional warped images based on their corresponding warp masks, helping prevent the model from overfitting to unreliable warping during training. **(3) Spatial-Temporal Backbone.** We introduce a lightweight spatio-temporal transformer to enforce cross-view and cross-frame coherence by incorporating frame-time embeddings and applying spatio-temporal attention. Together, these components form a view-conditional inpainting model that maintains stability in dynamic scenes while preserving the simplicity of pose-free supervision.

3.2.2 Realistic Warp Synthesis. Existing methods conditioned on warped images targets with random blocks [52], use 2D trackers for motion estimation [32], or ignore depth noise artifacts [92]. These heuristics miss two key artifacts in dynamic video: *camera motion* that pushes content off the frame and *object motion* that causes irregular object gaps. In addition, noisy monocular depths also add tearing patterns that simple masks cannot capture.

A perfect supervision signal would attach scale-aligned depths and ground-truth poses to every source-target video pair $(\mathbf{I}_s, \mathbf{I}_t)$, but collecting dense 6-DoF labels is prohibitively expensive. Instead, for each target frame, we procedurally create a warped image $\mathbf{I}_w$ and mask $\mathbf{M}$ that imitate test-time noise through four steps: *(1) Depth prediction.* We use a monocular estimator [83] to produce $\mathbf{D}_t$ and lift $\mathbf{I}_t$ into a point cloud. *(2) Random scene re-pose.* We segment the foreground objects using [54], designate the point corresponding to the center of the largest inscribed circle of the largest object as the world origin, then rotate the point cloud about a random axis by $[-30°, 30°]$, then translate by $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, (\lambda\bar{D})^2\mathbf{I}_3)$ ($\bar{D}$ is mean depth, λ is a scale factor, and $\mathbf{I}_3$ is the 3×3 identity matrix), yielding a synthetic transform $\mathbf{T}_w$. *(3) Forward projection.* We render the transformed point cloud with Z-buffer and bilinear splatting to preserve correct visibility, obtaining a warped image $\bar{\mathbf{I}}_t$ and depth $\bar{\mathbf{D}}_t$, this introduces tearing and stretching artifacts characteristic of real warps. *(4) Back-projection with noise.* We warp reversely with $\bar{\mathbf{I}}_t$ and $\bar{\mathbf{D}}_t$. As monocular depths are noisy, we perturb the inverse

transform $\mathbf{T}_w^{-1}$ with pose jitter $\delta\mathbf{T}$ to produce a slightly misaligned yet informative warp $\mathbf{I}_w$ that mirrors test-time noise. The same Z-buffer yields the binary mask $\mathbf{M}$, whose ragged contours precisely delineate disocclusions at borders and around moving objects, which is far richer than that in earlier work.

3.2.3 Noise-Adaptive Condition. Warp-conditioned diffusion models limit signal leakage by injecting step-dependent noise. Excessive noise degrades camera control, whereas insufficient noise encourages the network to replicate the warp. Dynamic inputs intensify this trade-off because depth errors are larger and data diversity is limited. To curb over-reliance on the warp latent, we apply a stronger perturbation.

$$x_\tau = \sqrt{\bar{\alpha}_{\tau/3}}\, z_w + \sqrt{1-\bar{\alpha}_{\tau/3}}\, \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, I), \tag{5}$$

where z_w is the latent encoding of the warped image, and x_τ denotes its noisy counterpart, the reduced index $\tau/3$ yields higher variance than that used for the target latent at the same step.

The informativeness of the warp condition directly impacts inpainting quality, unreliable warps should be down-weighted, while accurate ones should be preserved. We approximate warp fidelity by the mask density $\beta = \|\mathbf{M}\|_0/(BNHW)$, where $\|\mathbf{M}\|_0$ counts nonzero entries, B is the batch size, N the number of frames, and $H{\times}W$ the spatial resolution. A larger β indicates greater pixel overlap and thus higher reliability. We therefore modulate the blend between x_τ and the running target latent z_τ using the mask density

$$c_\tau = \left[\gamma\beta\, w_\tau x_\tau + (1-\gamma\beta\, w_\tau) z_\tau;\, \mathbf{M}\right], \tag{6}$$

where $w_\tau \in [0,1]$ decays monotonically with τ, and γ is a scale factor. Hence, reliable warps transmit more signal, whereas sparse or noisy warps are increasingly suppressed. Guided by the time-dependent weight w_τ, early iterations lock coarse geometry from x_τ, while later iterations refine details through z_τ, yielding effective error correction without sacrificing camera control.

3.2.4 Spatial–Temporal Backbone. Our architecture builds on the multi-view UNet of previous work [52] and introduces two compact temporal mechanisms. *(1) Frame-time embedding.* A sinusoidal encoding of the frame index is injected into every residual path, giving the feature hierarchy an explicit notion of temporal order. *(2) Spatial–temporal attention.* Each 2D transformer block is upgraded to a module that attends simultaneously to spatial tokens and their counterparts in neighboring frames, enforcing cross-frame coherence. These additions provide the temporal context needed for consistent geometry and appearance throughout the sequence.

Our view-conditional inpainting network eliminates the need for per-frame pose annotations by conditioning on depth-warped images, thereby leveraging rich appearance cues without explicit calibration. Coupled with realistic warp synthesis, noise-adaptive conditioning, and the proposed spatial–temporal backbone, this design yields a lightweight yet temporally consistent model that generates high-quality novel views from pose-free, dynamic video clips.

3.3 Spatial-Temporal Auto-Regressive Inference

At inference time, our system needs to convert a monocular dynamic sequence captured at a single view into a set of temporally aligned videos at multiple, user-specified fixed target viewpoints. These synthesized clips are later fused by downstream 4D pipelines [78] into a dense space-time representation. This task is limited by two factors. *(1) Viewpoint shifts.* Even moderate angular gaps can significantly amplify depth inaccuracies, directly warping from the source pose to a distant camera often produces tearing and occlusion holes, resulting in an unreliable conditioning signal for the inpainting stage. *(2) Long durations.* The diffusion backbone can denoise only a few dozen frames in a single forward pass, so naïvely processing a minute-long video breaks temporal coherence at window boundaries. We tackle both issues with an auto-regressive inference pipeline that proceeds **progressively in space**, advancing smoothly from the source view to the target view through a chain of virtual cameras, and **progressively in time**, sliding a partially overlapping window across the sequence. This dual recursion allows us to better handle viewpoint shifts and extended sequences while preserving consistent geometry and appearance throughout the generated views.

Fig. 3. **Overview of our spatiotemporal auto-regressive inference pipeline**. We decompose a single warp into small spline-interpolated hops. We then prepend the denoised clean latents to subsequent noisy windows, producing seamless long videos without extra memory overhead.

3.3.1 Spatial Auto-Regressive Expansion. A direct warp from the source view to a distant target pose often suffers from substantial depth noise and severe occlusions, which provide weak and unreliable cues for inpainting. To mitigate this issue, we decompose the overall viewpoint shift into a chain of smaller, well-conditioned hops that are easier to process. Following the design of ViewCrafter [93], we spline-interpolate a sequence of P virtual cameras that smoothly bridge the source and target views, ensuring that the transition remains gradual rather than abrupt. Starting from the source pose, we iterate through this chain of intermediate cameras: at each step, the current video is warped using the estimated depth to the next virtual pose, the resulting occlusion holes are inpainted, and a new depth map is re-estimated and scale-aligned with the previous step using the global solver of [52, 73]. The refined video produced at step p then becomes the input for step $p{+}1$. Because every hop spans only a few degrees, the magnitude of projection errors and the size of missing regions remain modest, which makes it easier for the diffusion model to thoroughly repair artifacts before proceeding. After P iterations, the sequence converges at the target view, producing coherent and high-quality frames. This gradual spatial expansion allows the system to robustly accommodate viewpoint

Table 1. **Comparative study with SoTA methods for 4D scene reconstruction** on the five different scenes from the iPhone [19] dataset. Symbols "↑" / "↓" denote scores that are the "higher" / "lower" the better.

Method	Venue	Apple	Block	Paper	Spin	Teddy	Avg
Metric: PSNR ↑							
GCD [69]	ECCV'24	9.82	12.30	9.75	10.37	11.61	10.77
ViewCrafter [93]	arXiv'24	10.19	10.28	10.63	11.15	11.50	10.75
Shape-of-Motion [74]	arXiv'24	11.06	11.72	11.93	11.28	10.42	11.28
DaS [22]	Siggraph'25	10.02	11.64	10.27	11.11	11.82	10.97
ReCamMaster [2]	arXiv'25	10.96	12.67	11.88	12.25	12.37	12.02
TrajectoryCrafter [92]	arXiv'25	13.88	14.21	14.89	14.51	13.73	14.24
See4D	**Ours**	**13.98**	**14.67**	**15.24**	**14.72**	**14.22**	**14.56**
Metric: SSIM ↑							
GCD [69]	ECCV'24	0.215	0.458	0.398	0.324	0.385	0.356
ViewCrafter [93]	arXiv'24	0.245	0.427	0.344	0.308	0.372	0.339
Shape-of-Motion [74]	arXiv'24	0.197	0.446	0.425	0.319	0.357	0.349
DaS [22]	Siggraph'25	0.217	0.388	0.356	0.312	0.381	0.331
ReCamMaster [2]	arXiv'25	0.264	0.454	0.471	0.344	0.401	0.387
TrajectoryCrafter [92]	arXiv'25	0.285	0.528	0.482	0.380	0.411	0.417
See4D	**Ours**	**0.309**	**0.555**	**0.514**	**0.399**	**0.434**	**0.442**
Metric: LPIPS ↓							
GCD [69]	ECCV'24	0.738	0.590	0.535	0.576	0.629	0.614
ViewCrafter [93]	arXiv'24	0.750	0.615	0.521	0.533	0.606	0.605
Shape-of-Motion [74]	arXiv'24	0.879	0.601	0.486	0.560	0.650	0.635
DaS [22]	Siggraph'25	0.732	0.593	0.520	0.551	0.608	0.601
ReCamMaster [2]	arXiv'25	0.683	0.537	0.491	0.545	0.572	0.566
TrajectoryCrafter [92]	arXiv'25	0.612	0.479	0.471	0.518	0.513	0.519
See4D	**Ours**	**0.581**	**0.455**	**0.439**	**0.501**	**0.486**	**0.492**

shifts while maintaining both appearance fidelity and geometric consistency throughout the traversal.

3.3.2 Temporal Auto-Regressive Inference. Hardware memory limits restrict each diffusion pass to processing only a relatively short clip, whereas the final output is typically required to remain coherent over much longer video horizons. To address this discrepancy, we introduce a sliding-window scheme for temporal autoregression. Let L denote the window length and $m < L$ the overlap size. We begin by denoising the first L frames to obtain an initial clean sequence. For each subsequent window, we concatenate the last m clean latent frames from the previous output with the next $L - m$ noisy frames from the current sequence, thereby forming an L-frame batch for denoising. The shared m-frame overlap serves as a temporal anchor, enforcing consistency in both appearance and motion, and effectively eliminating seam artifacts at window boundaries. Iterating this process over the entire sequence yields long videos that remain temporally smooth and coherent without increasing memory usage. In this way, spatial expansion decomposes viewpoint shifts into manageable steps, while temporal recursion amortizes memory costs and propagates dynamics across windows. Combined, these two strategies enable our pipeline to generate temporally consistent novel view videos across both viewpoint changes and extended sequences.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets. We train on two synthetic multi-view dynamic datasets with synchronized captures for 4D consistency. *SynCamMaster* [3] provides 34K clips from 3.4K scenes filmed by ten static cameras. *ReCamMaster* [2] adds motion, with 13.6K scenes captured by nine moving and one fixed camera, yielding 136K clips. Altogether, 170K temporally aligned videos cover diverse baselines, dynamics, environments, and trajectories. During training, we sample two clips (source/target) per SynCamMaster scene and the moving plus static clip per ReCamMaster scene.

Table 2. **Comparative study with SoTA methods for video generation** following VBench [33] evaluation protocol. The metrics measure video and frame qualities and consistencies. All scores are the higher the better.

Method	Subj. Consist	Back. Consist	Temp. Flick	Motion Smooth	Image. Quality	Aesth. Quality
DaS [22]	89.44	91.69	96.11	95.58	50.64	37.49
ReCamMaster [2]	90.56	93.42	95.11	**98.12**	52.47	39.65
TrajectoryCrafter [92]	89.61	92.55	92.78	93.49	52.20	37.62
See4D	**92.18**	**94.63**	**96.66**	97.87	**53.15**	**41.35**

Fig. 4. **Qualitative comparisons** of See4D and baselines for 4D reconstruction on the iPhone [19] dataset.

4.1.2 Implementation Details. We initialize from See3D weights [52]. Training uses 512×512 resolution, 16-frame sequences, batch size 16, and 10K iterations at 1×10^{-5} learning rate. We employ bfloat16 precision and classifier-free guidance [27] by dropping visual conditions with 0.1 probability. Training runs on eight NVIDIA A800-80GB GPUs for approximately 48 hours. At inference, we use a DDIM sampler [67] with the same guidance strategy.

4.1.3 Evaluation Protocol. We evaluate our method with two primary settings: 4D reconstruction and cross-view video generation. **4D Reconstruction.** We assess monocular 4D reconstruction on the public iPhone dataset [19]. Following [92], we select five handheld sequences, where the moving video serves as the source input, and the first synchronized fixed-camera clip as the target. Each scene

Fig. 5. **Qualitative comparisons** of TrajectoryCrafter [36] and See4D for 4D generation. A single frame from each example is shown due to space limits. Kindly refer to our **Webpage** at `see-4d.github.io` or **Appendix File** for high-resolution demo videos and frame-by-frame examples.

is furnished with COLMAP-derived poses and aligned depth maps from [73, 92], enabling rigorous geometric evaluation. Reconstruction quality is measured by PSNR, SSIM, and LPIPS [96].

Cross-View Video Generation. To more thoroughly evaluate our model's ability to synthesize a fixed target view from a single hand-held clip, we assemble an in-the-wild dataset of 200 monocular videos drawn from WebVid [4]. For each clip, we randomly select a static camera viewpoint as the target, spanning pan, tilt, and hemispherical arc trajectories around the actor's x- and y-axes. Quality is assessed using the VBench protocol [33].

Comparison Methods. We compare against serval state-of-the-art methods: GCD [69], ViewCrafter [93], DaS [22], TrajectoryCrafter [92], ReCamMaster [2], and Shape-of-Motion [73]. The first five are diffusion-based generative models, while Shape-of-Motion is the leading reconstruction-oriented 4D approach. Since ReCamMaster and DaS do not target reconstruction objectives, their reconstruction metrics are reported only for qualitative reference.

Table 3. **Ablation studies of proposed components** on iPhone [19]. Symbols "↑" / "↓" denote scores that are the "higher" / "lower" the better.

Module	Component	PSNR ↑	SSIM ↑	LPIPS ↓
Warp Syn. (*cf.* Sec.3.2.2)	from See3D [52]	11.93	0.377	0.568
	from Vivid4D [32]	12.23	0.391	0.556
	from TrajectoryCrafter [92]	12.48	0.420	0.537
	Ours	14.56	0.442	0.492
Noise Ada. (*cf.* Sec.3.2.3)	w/o noise adaptive condition	13.47	0.411	0.508
	Ours	14.56	0.442	0.492
Spa.-Tem. (*cf.* Sec.3.2.4)	w/o spatial-temporal transformer	10.66	0.347	0.685
	Ours	14.56	0.442	0.492
Auto.-Reg. (*cf.* Sec.3.3)	w/o spatial regression	13.58	0.424	0.518
	w/o temporal regression	13.02	0.438	0.517
	Ours	14.56	0.442	0.492

4.2 Comparative Study

4.2.1 4D Reconstruction. Tab. 1 shows that See4D attains the best average PSNR, SSIM, and LPIPS on all five iPhone scenes, with consistent gains per sequence, showing that our *trajectory-to-camera* formulation handles both rigid and deformable content. ReCamMaster relies on explicit pose input and fails outside its training frame, whereas our depth-warp conditioning needs no poses. ViewCrafter, designed for static images, loses 4D coherence when objects move, while our spatial–temporal transformer preserves alignment across time. TrajectoryCrafter and DaS assume clean geometric cues and degrade under noise, a weakness mitigated by our noise-adaptive condition. Shape-of-Motion leaves large voids when points disappear, but our warp synthesis and inpainting fill such occlusions.

4.2.2 Cross-View Video Generation. We evaluate on VBench's generation protocol [33] in Tab. 2, where baselines leverage large, dedicated video-diffusion backbones. Despite our core network being a multi-view synthesis model, See4D leads on five of six metrics and ranks a close second on motion smoothness. These results underscore the power of our view-conditional inpainting model and spatial–temporal auto-regressive inference pipeline. By fusing explicit geometry cues with robust temporal priors, our method exceeds the performance of models tailored solely for video generation.

Fig. 6. **Illustrative examples** of See4D for various applications (robotics, driving, games, movies, *etc.*). A single frame from each example is shown due to space limits. Kindly refer to our **Webpage** at `see-4d.github.io` or **Appendix File** for high-resolution demo videos and frame-by-frame examples.

4.2.3 Qualitative Assessment. Qualitative comparisons in Fig. 4 on iPhone reconstruction show that our method recovers sharp geometry and stable parallax with less flicker, outperforming Shape-of-Motion's blurring and TrajectoryCrafter's minor inconsistencies. Fig. 5 on cross-view generation demonstrates smooth, occlusion-aware texture synthesis and seamless transitions, while TrajectoryCrafter exhibits bleed-through and jitter. Additional cross-view generation examples are provided in Fig. 7. Please refer to our supplementary materials for more results and video demonstrations.

4.3 Ablation Study

Tab. 3 confirms that each component of our design is essential. Substituting our *realistic warp synthesis* with the simpler schemes of prior work causes a substantial drop in all reconstruction metrics, demonstrating that modeling real-world occlusions and depth noise is critical to provide a reliable view prior. Omitting *noise-adaptive condition* degrades performance across the board, showing that dynamically scaling perturbation to warp quality is key to preventing over-reliance on imperfect geometric cues and to encouraging robust inpainting. Removing the *spatial–temporal backbone* reduces reconstruction fidelity dramatically, underscoring the necessity of joint cross-frame attention for preserving coherent appearance in dynamic video content. Ablating either the spatial expansion or the temporal recursion in our *auto-regressive inference* pipeline leads to noticeable quality declines, illustrating that both progressive view hops and overlapping diffusion windows are indispensable for high-quality novel-view synthesis.

4.4 Downstream Applications

Beyond 4D reconstruction, See4D facilitates a range of downstream applications by producing temporally aligned, viewpoint-varying video sequences from a single input. These augmented videos with baseline variations enhance geometric understanding while maintaining appearance and motion consistency, making them suitable for diverse perception and creative tasks. In *robotic manipulation*, our method generates short turn-around sequences of handheld objects (e.g., wrenches or mugs), offering complementary perspectives for grasp planners to assess surface geometry and approach directions. The synthesized frames can enrich grasp simulation and contact evaluation without requiring multi-camera capture, supporting data-efficient policy learning and motion refinement. In *autonomous driving*, our method augments dash-cam recordings with different perspectives that reconstruct the surrounding context, including nearby vehicles and static structures. These additional views can improve visual coverage for perception and tracking modules, or serve as synthetic supervision for 3D detection under limited sensor setups. In *interactive gaming and virtual environments*, single-view gameplay recordings can be transformed into short multi-angle "fly-around" sequences that preserve temporal coherence and object dynamics, enabling more immersive replay or automatic camera switching. Similarly, in *cinematic post-production*, our method allows off-axis re-framing or stabilization from handheld footage, filling occlusions and maintaining motion continuity without requiring reshoots. Overall, our method demonstrates that viewpoint synthesis from monocular input can bridge perception and generation tasks, offering a lightweight, geometry-aware tool for robotic, autonomous, and creative systems.

5 Conclusion

In this work, we introduce See4D, a pose-free framework that converts a single hand-held video into a bank of synchronized, fixed-view sequences through a trajectory-to-camera formulation. Conditioning a latent diffusion model on depth-warped images and masks, our method removes the need for explicit 6-DoF annotations while preserving geometric cues. Unlike trajectory-to-trajectory methods, the proposed trajectory-to-camera formulation decouples scene modeling from camera motion, reducing training complexity and yielding more stable synthesis. A view-conditional inpainting model enforces cross-view and cross-frame consistency, while a spatial–temporal auto-regressive pipeline extends generation across viewpoints and long sequences, producing coherent videos that can be fused into dense 4D modeling and representation. Experiments show improved cross-view quality and reconstruction accuracy over existing baselines, and we demonstrate benefits across applications such as robotic manipulation, autonomous driving, interactive gaming, and cinematic post-production.

Fig. 7. Generated examples of See4D on 4D generation. The source video is from KLing. Kindly refer to our **Webpage** at see-4d.github.io for high-resolution demo videos and frame-by-frame examples.

Fig. 8. Generated examples of See4D on 4D generation. The source videos are from Video Prediction Policy, Vista, and ReCamMaster. Kindly refer to our **Webpage** at see-4d.github.io for high-resolution demo videos and frame-by-frame examples.

Fig. 9. Generated examples (#1, #2, #3) of See4D on 4D generation. The source video is from KLing (Credit to: https://app.klingai.com).

Fig. 10. Generated examples (#4, #5, #6) of See4D on 4D generation. The source video is from KLing (Credit to: https://app.klingai.com).

Fig. 11. Generated examples (#7, #8, #9) of See4D on 4D generation. The source video is from KLing (Credit to: https://app.klingai.com).

Fig. 12. Application use cases of See4D on robot grasping. The source videos are from Video Prediction Policy (Credit to: https://video-prediction-policy.github.io).

Fig. 13. Application use cases of See4D on robot grasping. The source videos are from Video Prediction Policy (Credit to: https://video-prediction-policy.github.io).

Fig. 14. Application use cases of See4D on driving scene generation. The source videos are from Vista (Credit to: https://opendrivelab.com/Vista).

Fig. 15. Application use cases of See4D on driving scene generation. The source videos are from Vista (Credit to: https://opendrivelab.com/Vista).

Fig. 16. Application use cases of See4D on computer game generation. The source video is from YouTube (Credit to: https://www.youtube.com).

Fig. 17. Application use cases of See4D on movie clip generation. The source videos are from ReCamMaster (Credit to: https://jianhongbai.github.io/ReCamMaster).

Fig. 18. Qualitative comparisons of See4D and TrajectoryCrafter on video generation. The source video is from KLing (Credit to: https://app.klingai.com).

Fig. 19. Qualitative comparisons of See4D and TrajectoryCrafter on video generation. The source video is from KLing (Credit to: https://app.klingai.com).

Fig. 20. Qualitative comparisons of See4D and TrajectoryCrafter on video generation. The source video is from KLing (Credit to: https://app.klingai.com).

Fig. 21. Qualitative comparisons of See4D and TrajectoryCrafter on video generation. The source video is from KLing (Credit to: https://app.klingai.com).

References

[1] Christoph Anthes, Rubén Jesús García-Hernández, Markus Wiedemann, and Dieter Kranzlmüller. 2016. State of the art of virtual reality technology. In *2016 IEEE aerospace conference*. IEEE, 1–19.

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. 2025. ReCamMaster: Camera-Controlled Generative Rendering from A Single Video. *arXiv preprint arXiv:2503.11647* (2025).

[3] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. 2024. SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints. *arXiv preprint arXiv:2412.07760* (2024).

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1728–1738.

[5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*. 5855–5864.

[6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5470–5479.

[7] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2023. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *IEEE/CVF International Conference on Computer Vision*. 19697–19705.

[8] Weikang Bian, Zhaoyang Huang, Xiaoyu Shi, Yijin Li, Fu-Yun Wang, and Hongsheng Li. 2025. GS-DiT: Advancing Video Generation with Pseudo 4D Gaussian Fields through Efficient Dense 3D Point Tracking. *arXiv preprint arXiv:2501.02690* (2025).

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

[10] Grigore C Burdea and Philippe Coiffet. 2003. *Virtual reality technology*. John Wiley & Sons.

[11] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19457–19467.

[12] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. 2024. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055* (2024).

[13] Shen Chen, Jiale Zhou, and Lei Li. 2024. Optimizing 3D Gaussian Splatting for Sparse Viewpoint Scene Reconstruction. *arXiv preprint arXiv:2409.03213* (2024).

[14] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. 2024. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*. Springer, 370–386.

[15] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. 2021. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE Computer Society, 14304–14314.

[16] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 2024. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

[17] Fanda Fan, Chaoxu Guo, Litong Gong, Biao Wang, Tiezheng Ge, Yuning Jiang, Chunjie Luo, and Jianfeng Zhan. 2023. Hierarchical masked 3d diffusion model for video outpainting. In *ACM International Conference on Multimedia*. 7890–7900.

[18] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. 2024. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309* (2024).

[19] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. 2022. Monocular dynamic view synthesis: A reality check. In *Advances in Neural Information Processing Systems*, Vol. 35. 33768–33780.

[20] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. 2024. Cat3d: Create anything in 3d with multi-view diffusion models. In *Advances in Neural Information Processing Systems*, Vol. 37. 75468–75494.

[21] Bohai Gu, Hao Luo, Song Guo, and Peiran Dong. 2024. Advanced Video Inpainting Using Optical Flow-Guided Efficient Diffusion. *arXiv preprint arXiv:2412.00857* (2024).

[22] Zekai Gu, Rui Yan, Jiahao Lu, Peng Li, Zhiyang Dou, Chenyang Si, Zhen Dong, Qifeng Liu, Cheng Lin, Ziwei Liu, et al. 2025. Diffusion as Shader: 3D-aware Video Diffusion for Versatile Video Generation Control. *arXiv preprint arXiv:2501.03847* (2025).

[23] Antoine Guedon and Vincent Lepetit. 2024. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5354–5363.

[24] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *IEEE/CVF International Conference on Computer Vision*. 19740–19750.

[25] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. *arXiv preprint arXiv:2404.02101* (2024).

[26] Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. 2025. CameraCtrl II: Dynamic Scene Exploration via Camera-controlled Video Diffusion Models. *arXiv preprint arXiv:2503.10592* (2025).

[27] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[28] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. 2022. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868* (2022).

[29] Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. 2025. EX-4D: EXtreme Viewpoint 4D Video Synthesis via Depth Watertight Mesh. *arXiv preprint arXiv:2506.05554* (2025).

[30] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. 2024. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095* (2024).

[31] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2024. 2d Gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH Conference*. 1–11.

[32] Jiaxin Huang, Sheng Miao, BangBnag Yang, Yuewen Ma, and Yiyi Liao. 2025. Vivid4D: Improving 4D Reconstruction from Monocular Video by Video Inpainting. *arXiv preprint arXiv:2504.11092* (2025).

[33] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive Benchmark Suite for Video Generative Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.

[34] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. 2022. CenterSnap: Single-Shot Multi-Object 3D Shape Reconstruction and Categorical 6D Pose and Size Estimation. In *IEEE International Conference on Robotics and Automation*. 10632–10640.

[35] Hyeonho Jeong, Suhyeon Lee, and Jong Chul Ye. 2025. Reangle-A-Video: 4D Video Generation as Video-to-Video Translation. *arXiv preprint arXiv:2503.09151* (2025).

[36] Longbin Ji, Lei Zhong, Pengfei Wei, and Changjian Li. 2025. PoseTraj: Pose-Aware Trajectory Control in Video Diffusion. *arXiv preprint arXiv:2503.16068* (2025).

[37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.

[38] Diederik P Kingma, Max Welling, et al. 2013. Auto-encoding variational bayes.

[39] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. 2017. Tanks and Temples: Benchmarking Large-Scale Scene Reconstruction. *ACM Transactions on Graphics* 36, 4 (2017), 1–13.

[40] Steven M LaValle. 2023. *Virtual reality*. Cambridge university press.

[41] Minhyeok Lee, Suhwan Cho, Chajin Shin, Jungho Lee, Sunghun Yang, and Sangyoun Lee. 2025. Video diffusion models are strong video inpainter. In *AAAI Conference on Artificial Intelligence*. 4526–4533.

[42] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. 2024. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20775–20785.

[43] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. 2025. DiffuEraser: A Diffusion Model for Video Inpainting. *arXiv preprint arXiv:2501.10018* (2025).

[44] Yaowei Li, Lingen Li, Zhaoyang Zhang, Xiaoyu Li, Guangzhi Wang, Hongxiang Li, Xiaodong Cun, Ying Shan, and Yuexian Zou. 2025. BlobCtrl: A Unified and Flexible Framework for Element-level Image Generation and Editing. *arXiv preprint arXiv:2503.13434* (2025).

[45] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 2021. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6498–6508.

[46] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. 2023. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4273–4284.

[47] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. 2024. ReconX: Reconstruct Any Scene from

Sparse Views with Video Diffusion Model. *arXiv preprint arXiv:2408.16767* (2024).

[48] Kunhao Liu, Ling Shao, and Shijian Lu. 2024. Novel View Extrapolation with Video Diffusion Priors. *arXiv preprint arXiv:2411.14208* (2024).

[49] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. 2023. Zero-1-to-3: Zero-shot one image to 3d object. In *IEEE/CVF International Conference on Computer Vision*. 9298–9309.

[50] Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. 2025. Free4D: Tuning-free 4D Scene Generation with Spatial-Temporal Consistency. *arXiv preprint arXiv:2503.20785* (2025).

[51] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. 2024. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20654–20664.

[52] Baorui Ma, Huachen Gao, Haoge Deng, Zhengxiong Luo, Tiejun Huang, Lulu Tang, and Xinlong Wang. 2025. You See it, You Got it: Learning 3D Creation on Pose-Free Videos at Scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[53] Yue Ma, Kunyu Feng, Xinhua Zhang, Hongyu Liu, David Junhao Zhang, Jinbo Xing, Yinhan Zhang, Ayden Yang, Zeyu Wang, and Qifeng Chen. 2025. Follow-Your-Creation: Empowering 4D Creation through Video Inpainting. *arXiv preprint arXiv:2506.04590* (2025).

[54] Maxwell Meyer and Jack Spruyt. 2025. BEN: Using Confidence-Guided Matting for Dichotomous Image Segmentation. *arXiv preprint arXiv:2501.06230* (2025).

[55] Sheng Miao, Jiaxin Huang, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Andreas Geiger, and Yiyi Liao. 2024. Efficient Depth-Guided Urban View Synthesis. *arXiv preprint arXiv:2407.12395* (2024).

[56] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5480–5490.

[57] Byeongjun Park, Hyojun Go, Hyelin Nam, Byung-Hoon Kim, Hyungjin Chung, and Changick Kim. 2025. SteerX: Creating Any Camera-Free 3D and 4D Scenes with Geometric Steering. *arXiv preprint arXiv:2503.12024* (2025).

[58] Jangho Park, Taesung Kwon, and Jong Chul Ye. 2025. Zero4D: Training-Free 4D Video Generation From Single Video Using Off-the-Shelf Video Diffusion Model. *arXiv preprint arXiv:2503.22622* (2025).

[59] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable Neural Radiance Fields. In *IEEE/CVF International Conference on Computer Vision*. 5865–5874.

[60] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. 2021. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228* (2021).

[61] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988* (2022).

[62] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10318–10327.

[63] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Muller, Alexander Keller, Sanja Fidler, and Jun Gao. 2025. Gen3c: 3d-informed world-consistent video generation with precise camera control. *arXiv preprint arXiv:2503.03751* (2025).

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

[65] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. 2023. Zeroshot 360-degree view synthesis from a single real image. *arXiv preprint arXiv:2310.17994* (2023).

[66] Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. 2024. Genwarp: Single image to novel views with semantic-preserving generative warping. In *Advances in Neural Information Processing Systems*, Vol. 37. 80220–80243.

[67] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

[68] Stanislaw Szymanowicz, Jason Y. Zhang, Pratul Srinivasan, Ruiqi Gao, Arthur Brussee, Aleksander Holynski, Ricardo Martin-Brualla, Jonathan T. Barron, and Philipp Henzler. 2025. Bolt3D: Generating 3D Scenes in Seconds. *arXiv preprint arXiv:2503.14445* (2025).

[69] Basile Van Hoorick, Rundi Wu, Ege Ozguroglu, Kyle Sargent, Ruoshi Liu, Pavel Tokmakov, Achal Dave, Changxi Zheng, and Carl Vondrick. 2024. Generative camera dolly: Extreme monocular dynamic novel view synthesis. In *European Conference on Computer Vision*. Springer, 313–331.

[70] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. 2021. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994* (2021).

[71] Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. 2024. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. In *European Conference on Computer Vision*. Springer, 153–168.

[72] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. 2023. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *IEEE/CVF International Conference on Computer Vision*. 9065–9076.

[73] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. 2024. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764* (2024).

[74] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. 2024. Shape of Motion: 4D Reconstruction from a Single Video. *arXiv preprint arXiv:2407.13764* (2024).

[75] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20697–20709.

[76] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2021. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021).

[77] Isabell Wohlgenannt, Alexander Simons, and Stefan Stieglitz. 2020. Virtual reality. *Business & Information Systems Engineering* 62 (2020), 455–461.

[78] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20310–20320.

[79] Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. 2024. Cat4d: Create anything in 4d with multi-view video diffusion models. *arXiv preprint arXiv:2411.18613* (2024).

[80] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. 2024. Reconfusion: 3d reconstruction with diffusion priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21551–21561.

[81] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. 2021. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9421–9431.

[82] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. 2024. Street Gaussians for Modeling Dynamic Urban Scenes. *arXiv preprint arXiv:2401.01339* (2024).

[83] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. 2024. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815* (2024).

[84] Jiawei Yang, Marco Pavone, and Yue Wang. 2023. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8254–8263.

[85] Ling Yang, Kaixin Zhu, Juanxi Tian, Bohan Zeng, Mingbao Lin, Hongjuan Pei, Wentao Zhang, and Shuicheng Yan. 2025. WideRange4D: Enabling High-Quality 4D Reconstruction with Wide-Range Movements and Scenes. *arXiv preprint arXiv:2503.13435* (2025).

[86] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. 2024. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 20331–20341.

[87] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072* (2024).

[88] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. 2023. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642* (2023).

[89] Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. 2025. SV4D 2.0: Enhancing Spatio-Temporal Consistency in Multi-View Video Diffusion for High-Quality 4D Generation. *arXiv preprint arXiv:2503.16396* (2025).

[90] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. 2020. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5336–5345.

[91] Hanyang Yu, Xiaoxiao Long, and Ping Tan. 2024. LM-Gaussian: Boost Sparse-view 3D Gaussian Splatting with Large Model Priors. *arXiv preprint arXiv:2409.03456* (2024).

[92] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. 2025. TrajectoryCrafter: Redirecting Camera Trajectory for Monocular Videos via Diffusion Models. *arXiv preprint arXiv:2503.05638* (2025).

[93] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. 2024. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048* (2024).

[94] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. 2024. Mip-splatting: Alias-free 3d gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19447–19456.

[95] Jiawei Zhang, Jiahe Li, Xiaohan Yu, Lei Huang, Lin Gu, Jin Zheng, and Xiao Bai. 2024. CoR-GS: Sparse-View 3D Gaussian Splatting via Co-Regularization. In *European Conference on Computer Vision*. Springer, 335–352.

[96] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[97] Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang, Yinan Zhao, Peter Vajda, Dimitris Metaxas, and Licheng Yu. 2024. Avid: Any-length video inpainting with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7162–7172.

[98] Sijie Zhao, Wenbo Hu, Xiaodong Cun, Yong Zhang, Xiaoyu Li, Zhe Kong, Xiangjun Gao, Muyao Niu, and Ying Shan. 2024. Stereocrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos. *arXiv preprint arXiv:2409.07447* (2024).

[99] Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. 2025. GenXD: Generating any 3D and 4D scenes. In *International Conference on Learning Representations*.

[100] JM Zheng, KW Chan, and Ian Gibson. 1998. Virtual reality. *Ieee Potentials* 17, 2 (1998), 20–23.

[101] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. 2024. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21336–21345.

[102] Zhizhuo Zhou and Shubham Tulsiani. 2023. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12588–12597.

[103] Bojia Zi, Shihao Zhao, Xianbiao Qi, Jianan Wang, Yukai Shi, Qianyu Chen, Bin Liang, Kam-Fai Wong, and Lei Zhang. 2025. Cococo: Improving text-guided video inpainting for better consistency, controllability and compatibility. In *AAAI Conference on Artificial Intelligence*. 11067–11076.