Unified Panoramic–Gaussian Representation for Monocular 4D Scene Synthesis

Yuankun Yang[1], Yi Wei[2], Wenyang Zhou[2], and Li Zhang[1†]

[1]School of Data Science, Fudan University
[2]Central Media Technology Institute, Huawei

github.com/LogosRoboticsGroup/PanoGaussian

Abstract. 4D scene synthesis from monocular videos has made significant progress in recent years. However, existing methods are typically constrained by view interpolation. As a result, they struggle to infer unseen regions beyond the observed views. In this paper, we reformulate the task as 4D scene synthesis with unseen regions, which extends beyond traditional interpolation settings. Camera-conditioned video generation enables unseen region synthesis by guiding generation along specified cameras. However, these methods lack explicit 3D priors and are optimized with random camera trajectories. This design leads to severe inconsistencies under large trajectory deviations. To address this limitation, we build a unified training and inference framework with panoramic trajectory guidance. While this design improves cross-view consistency, the panoramic representation alone fails to model dynamic content effectively. Object motion in panoramic space introduces scale and shape distortions. To address this, we propose PanoGaussian, a unified Panoramic-Gaussian representation that distills the panoramic representation into an explicit dynamic Gaussian representation to capture dynamic physical priors of the 4D scene. Experiments demonstrate that PanoGaussian achieves consistent 4D scene synthesis even under large viewpoint variations.

Keywords: Panoramic Representation · Dynamic Gaussian Splatting · Monocular 4D Scene Synthesis · Video Generation

1 Introduction

Reconstruction of 4D scenes from monocular videos holds considerable promise for immersive creation and virtual reality. Recent advances mainly focus on novel view synthesis. They are typically formulated as a view interpolation task, which interpolates viewpoints within the observed camera trajectory. However, this task cannot be extrapolated beyond the captured range of the camera, resulting in incomplete reconstruction in unseen regions.

† Corresponding author: lizhangfd@fudan.edu.cn

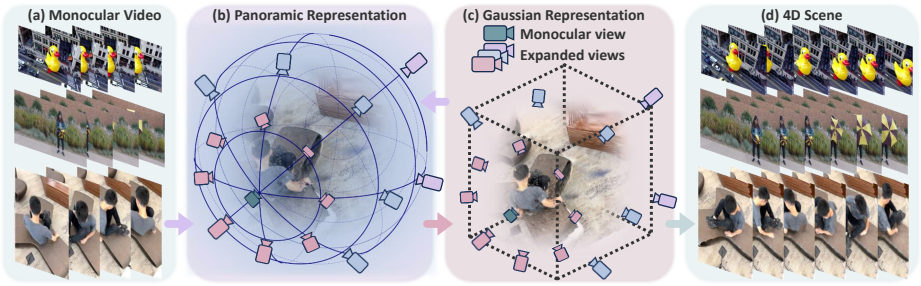

Fig. 1: PanoGaussian. (a) Given a monocular video, (b) PanoGaussian constructs a panoramic representation for scene exploration under a progressive strategy. (c) This panoramic representation is unified with Gaussian representation to achieve geometry-preserved dynamic scene modeling. (d) The unified Panoramic–Gaussian representation enables geometrically consistent 4D scene synthesis across unseen regions.

To address this limitation, we aim to reconstruct the entire 4D scene including both view interpolation and extension into unseen regions. A straightforward choice is to leverage video generation models as video generative priors to extend the unseen regions. However, video-level generative priors generally lack 3D prior knowledge. This leads to distortion in 3D geometry when rendering from novel viewpoints. Although some video generation models incorporate 3D consistency through depth re-projection or 3D datasets, they are generally optimized with small and random camera trajectories. As a result, these methods exhibit instability and are only effective for minor viewpoint changes. When extrapolated to large camera angles, they produce severe geometric distortions.

Therefore, we adopt a panoramic trajectory paradigm, which aims at narrowing the gap between training and inference of generative priors to ensure geometric stability. This paradigm designs a fixed-angle shift strategy for geometric reprojection when training the video generative prior, enabling the model to learn consistent synthesis. During inference, we leverage the panorama coordinates to iteratively extend the scene. Panorama coordinates provide a natural global context for full-scene coverage. Therefore, it ensures spatially uniform extension of scenes while requiring minimal reliance on generative priors. However, panoramic projections are ill-suited for modeling dynamic content. Object movements in panoramic space usually cause deformation in scale and shape. Consequently, an explicit 3D representation is still necessary to preserve geometry while capturing dynamics.

To this end, we propose PanoGaussian, which unifies the panoramic representation with dynamic Gaussian Splatting representation for geometry-preserved 4D scene modeling, as illustrated in Fig. 1. Specifically, PanoGaussian first leverages the panorama coordinates to design the camera trajectory for iterative view expansion. These expanded camera trajectories are processed by the video generative prior to synthesizing novel views. The generated results are subsequently distilled into the dynamic Gaussian Splatting representation for geometrically consistent 4D reconstruction.

Our main contributions are summarized as follows: 1) We propose PanoGaussian, a unified Panoramic-Gaussian representation that distills video generative priors into explicit 3D dynamic Gaussians through masked refinement, preserving geometry and motion under large viewpoint changes. 2) We design progressive expansion with fixed-geometry warping and region-wise supervision that structurally prevents error accumulation during unseen-region synthesis. 3) PanoGaussian achieves consistent 4D scene synthesis, maintaining geometric fidelity even under large viewpoint changes.

2 Related work

Dynamic Novel View Synthesis is commonly formulated as view interpolation from a set of posed images. Early dynamic reconstruction approaches explored neural scene representations (NeRFs) by treating time as an additional coordinate [14, 15, 17, 37, 58, 65, 72]. More recent efforts extended Gaussian Splatting (3DGS) to the 4D domain by modeling 3D Gaussian trajectories over time [10, 12, 29, 36, 44, 48, 67, 89], learning time-conditioned deformation networks [20, 41, 78, 85], or directly parameterizing Gaussians with 4D means and covariances [11, 84]. In monocular settings [8, 33, 76, 97], Gaussian Marbles [66] introduced isotropic Gaussian constraints to jointly model motion and appearance. MoSca [32] integrated 2D foundation model priors to guide motion scaffold optimization. USplat4D [19] modeled motion uncertainty to improve rendering robustness. Shape of Motion [75] exploited the low-dimensional structure of scene motion to regularize 3D trajectories. Recent feed-forward approaches [40, 42, 43, 50, 63, 77, 83] aimed to directly estimate 3DGS attributes to eliminate per-scene optimization. Despite these advances, all existing methods struggle to extrapolate beyond the observed camera range, yielding incomplete reconstructions in unseen regions.

Camera-Conditioned Video Generation aims to synthesize controllable videos along specified camera trajectories. Early approaches [4, 26] pioneered conditional video generation by conditioning diffusion architecture [25, 53, 60, 91] on the first-frame latent representation for temporal coherence. Building upon these foundation models, recent camera-conditioned video generation models [5, 6, 59, 70, 82, 86, 88, 90, 94] condition video diffusion on projected 3D point clouds derived from target camera trajectories for geometry-aware synthesis. In contrast, another related line of methods [2, 21, 22, 49, 95] discards explicit reprojection, instead encoding camera motion through learned or Plücker-based camera embeddings to facilitate flexible viewpoint control. However, these frameworks lack holistic 3D priors, often leading to spatial inconsistencies under large viewpoint changes. PanoGaussian addresses this limitation by injecting a unified Panoramic-Gaussian representation that enables complete 4D scene perception and enforces geometric consistency throughout the generated videos.

Panoramic Scene Synthesis captures a wide, continuous field of view for immersive scene representation. Traditional methods design panoramas as concentric mosaics [64], layered depth map [62, 93] or mesh representations [23, 24] for improved rendering of object surfaces. Recent panoramic scene reconstruction approaches optimize neural radiance fields [9, 18, 27, 46, 73] or 3D Gaussian Splatting [3, 7, 31, 34] directly from a single panorama to render static volume of the whole scene. These methods further advance the generative panorama outpainting [1, 35, 45, 47, 74, 81, 92] through the diffusion denoising process. However, directly extending panorama to 4D scene reconstruction from monocular videos remains challenging, as object motion in panoramic space often causes geometric distortion. To address this, PanoGaussian introduces a Panoramic-Gaussian representation that preserves geometry while capturing dynamic object motion.

3 Preliminaries

3D Gaussian Splatting explicitly represents a scene as a collection of 3D Gaussians [30]. Each Gaussian is defined by color $\mathbf{c} \in \mathbb{R}^3$, opacity $\alpha \in \mathbb{R}$, position mean $\mu \in \mathbb{R}^3$ and covariance matrix $\Sigma \in \mathbb{R}^{3\times3}$. However, Gaussian Splatting typically relies on pixel-level supervision during optimization, which restricts its ability to generalize beyond observed views.

Video Generative Priors enable controllable view synthesis by guiding video generation along a specified camera trajectory. Given an input video $\boldsymbol{x}_0 \in \mathbb{R}^{\mathrm{n}\times3\times\mathrm{h}\times\mathrm{w}}$, the forward diffusion process progressively injects Gaussian noise as $\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_0 + \sigma_t \epsilon$, where $\alpha_t \in (0, 1)$ and σ_t are the signal and noise scaling coefficients in timestep t. The reverse process learns to recover the clean video distribution by progressively denoising:

$$\min_{\theta} \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}[\|\epsilon_\theta(\boldsymbol{x}_t, t) - \epsilon\|_2^2], \tag{1}$$

where $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ is the network-predicted noise parameterized by θ, allowing progressive recovery of $\boldsymbol{x}_0$ from $\boldsymbol{x}_t$. Despite their strong generative ability, video diffusion usually lacks explicit 3D geometric understanding. They are effective for small viewpoint changes, but often fail under large camera motions.

Panoramic Representation captures a complete $360^\circ \times 180^\circ$ view of a scene. It is usually defined in a spherical domain. This representation is parameterized by the coordinates (ϕ, θ), where ϕ is the azimuth and θ the elevation. Although providing global scene coverage, the panoramic projection introduces geometric distortions. Object motion in panoramic space leads to scale and shape deformation. Therefore, bridging this gap between panoramic representation and 3D geometry is essential for geometry-aware 4D scene synthesis.

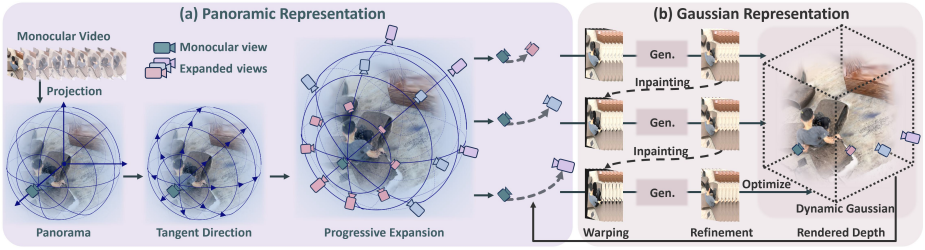

Fig. 2: Architecture Overview. (a) PanoGaussian begins by reprojecting the monocular video into panoramic coordinates for full-scene exploration (Sec. 4.1). Under this panorama, tangent directions are constructed to progressively guide trajectory expansion for unseen regions (Sec. 4.2). (b) PanoGaussian employs Gaussian-rendered depth to warp the original video frames along these expanded trajectories. The warped videos are then inpainted and refined by a video generative prior pretrained under panoramic trajectory paradigm. The refined results are distilled into Dynamic Gaussians for geometry-preserving 4D scene synthesis (Sec. 4.3).

4 Methods

Given a monocular video, PanoGaussian aims to extend conventional view interpolation to geometry-aware 4D scene reconstruction with coverage of unseen regions. Our core contribution is a unified Panoramic-Gaussian representation that distills panoramic trajectory-guided generative outputs into explicit 3D dynamic Gaussians, as illustrated in Fig. 2. To achieve this, we adopt a panoramic trajectory paradigm for camera-conditioned video priors (Sec. 4.1), progressively expand scene coverage under fixed-geometry warping (Sec. 4.2), and unify panoramic exploration with dynamic Gaussian optimization via masked refinement (Sec. 4.3).

4.1 Panoramic Trajectory Paradigm

To condition video generative priors on target viewpoints, we project observations into a panoramic domain and keep training and inference under the same warping protocol. Specifically, we apply a fixed panoramic trajectory protocol for both training and inference of the video generative priors.

Panoramic Trajectories Training. We design a fixed panoramic angle to guide the model in learning panoramic-aware motion and appearance generation during training. Specifically, for each source video $V_{src,tr}$ in the training set, we construct a panoramic trajectory following a random direction d_{tr} and a fixed panoramic elevation θ. This panoramic trajectory is generated by uniform sampling along the panoramic direction to simulate the camera motion. Following the exact configuration of previous projection-based generative priors [6], we conduct a double reprojection strategy along the defined trajectory to produce

a warped covisible video $V_{cov,tr}$. The masked video is then applied to train the video diffusion prior. It ensures that the model learns to synthesize content consistent with panoramic viewpoints. This process enables video generative priors to internalize panoramic spatial continuity.

Panoramic Paradigm Inference. In the inference stage, we design a progressive extension of the 4D scene within the panoramic coordinate system. Specifically, for each video frame $I_t \in V_{\text{src,inf}}$, we extract its corresponding depth map D_t and camera poses (rotation R_t, intrinsics K_t) from the optimized 4D Gaussian Splatting (4D GS) backbone. We unproject these pixels into a sequence of 3D point clouds $P_t = \Phi^{-1}([I_t, D_t], R_t, K_t)$ through inverse perspective projection. Crucially, we assume that depth D_t and poses $[R_t, K_t]$ derived from 4D GS are fixed during this step. Because this process inherently operates in 3D world-space coordinates, camera translation, dynamic motion, and parallax are naturally handled without requiring any image-space approximations. The center of the projected point cloud is defined as the origin of the panorama $P^c = (x^c, y^c, z^c)$. We then map the 3D points P_t into the panoramic space. The full projection implementation details are provided in the supplementary material. This ensures that the first camera direction corresponds to x-axis for uniform scene exploration.

In this way, we achieve both training and inference in correspondence under the panoramic paradigm. This alignment reduces domain gaps between optimization and generation of the generative prior. We then perform progressive view expansion (Sec. 4.2) within this aligned system.

4.2 Progressive Expansion Strategy

After mapping all video pixels into the panoramic coordinate system, PanoGaussian applies a progressive expansion strategy to gradually explore the full 4D scene. Naive or random view expansion often suffers from error accumulation and inconsistency between multiple trajectories. To address this inefficiency, we design a regularized expansion method that ensures comprehensive yet coherent scene exploration.

Panoramic Trajectory Construction. For view expansion, we use M evenly spaced directions on the panoramic sphere:

$$d_m = (\cos(\tfrac{2\pi m}{M}), \sin(\tfrac{2\pi m}{M})), \quad m = 0, \ldots, M-1, \tag{2}$$

where each d_m represents a tangent direction for view expansion. Along each direction d_m, the camera view is expanded E times with an angular step α. The accumulated rotation in the expansion step e is expressed as:

$$\mathbf{s}_{m,e} = \mathbf{n}\cos(e\alpha) + (\mathbf{a}_m \times \mathbf{n})\sin(e\alpha), \quad e = 0, \ldots, E, \tag{3}$$

where $\mathbf{n}$ is the initial camera viewing vector, and $\mathbf{a}_m$ denotes the unit rotation axis orthogonal to $\mathbf{n}$ in the tangent plane defined by d_m. This procedure produces

E uniformly interpolated panoramic trajectories that progressively cover the entire scene. Each trajectory contains T frames over time, denoted as $P_{e,t}$ where $e = 0, \ldots, E-1$ and $t = 0, \ldots, T-1$.

Progressive Scene Expansion. Following prior camera-controlled generation pipelines [6,90], we warp source frames onto each trajectory and inpaint masked regions with the video diffusion prior. For each target video frame $I_t \in V_{\text{src,inf}}$, we first warp it onto the panoramic trajectory $P_{e,t}$, resulting in the warped $I_{e,t}$. These warped frames $\{I_{e,t}\}_{t=0}^{T-1}$ form the intermediate video $V_{e,\text{warped}}$. We then apply progressive generation with the video diffusion prior to the panoramic representation. Specifically, the warped intermediate video $V_{e,\text{warped}}$ and a binary mask indicating the unobserved regions are directly provided as conditioning inputs for the video diffusion to perform inpainting. At every step, warping uses the original observed video $V_{\text{src,inf}}$ with fixed 4D GS geometry rather than previously hallucinated outputs, and the mask confines supervision to newly generated regions, structurally preventing error accumulation across steps. Each warped video $V_{e,\text{warped}}$ is successively refined to produce $V_{e,\text{refined}}$. This procedure extends the view and recovers details in unobserved regions. After each refinement step, the updated video $V_{e,\text{refined}}$ is warped to the next unrefined trajectory $V_{e+1,\text{warped}}$ to fill the missing content, while preserving areas that have already warped. This iterative process continues until all panoramic trajectories are refined, resulting in videos that are spatially continuous and mutually consistent across all directions. The refined videos are then used to supervise the Panoramic-Gaussian representation in Sec. 4.3.

4.3 Panoramic-Gaussian Representation

The panoramic coordinates provide a natural global context for capturing the entire scene. However, object motion in panoramic space often leads to deformation in scale and shape. This deformation disrupts the underlying physical dynamics, resulting in inconsistent geometry and unrealistic motion. To address this, PanoGaussian introduces a composite Panoramic-Gaussian representation. This design leverages panoramic coverage to provide global consistency while applying Gaussian priors to preserve geometry.

Unified Gaussian Geometry. We first optimize dynamic Gaussian Splatting on reference videos to obtain a stable Gaussian representation of the observed scene. From this representation, we extract panoramic trajectories $P_{e,t}$ as described in Sec. 4.2, and render $V_{e,\text{render}}$ based on these camera trajectories from the Gaussian representation. Previous camera-controlled video generation methods [59,90] estimate depth independently for each frame. This operation introduces large depth estimation errors, causing severe inconsistency and broken geometry. In contrast, our unified Panoramic-Gaussian design directly projects the stable depth maps from known views to novel viewpoints as geometric references. Unseen regions are seamlessly inpainted by the generative model without

requiring pre-existing depth, and their newly hallucinated geometry is subsequently distilled back to update the 4D Gaussian model. This progressive depth initialization ensures consistent geometric alignment and prevents structure collisions during panoramic rendering.

Refinement Regulation. After obtaining the refined video $V_{e,\text{refined}}$, we introduce a refinement regulation to enforce pixel-wise consistency. Specifically, we design a masked mean squared error (MSE) between Gaussian rendering and panoramic-trajectory refined videos:

$$L_{\text{refine}} = \text{MSE}(M_e \cdot V_{e,\text{refined}}, M_e \cdot V_{e,\text{render}}). \tag{4}$$

Here, M_e is a binary mask indicating the unknown regions in the warped video $V_{e,\text{warped}}$. It is generated by identifying pixels with no projected points from the known views. To ensure consistency across the refinement process, M_e is kept fixed during the current expansion step e, focusing the loss strictly on the newly hallucinated content. This refinement is jointly optimized with the source video $V_{\text{src,inf}}$, guided by all geometric regularization in MoSca [32]. We update the progressive scene expansion after specific Gaussian optimization iterations. Specifically, new Gaussians are initialized from the inpainted regions and optimized via the MSE loss against the refined video, without requiring explicit depth estimates from the diffusion model. This could produce improved Gaussian geometry for panoramic warping and generate updated panoramic trajectories to supervise the dynamic Gaussian.

Through this unified optimization, each video pixel along the panoramic trajectory directly corresponds to its Gaussian-rendered counterpart. This correspondence forms a pixel-to-pixel alignment between panoramic and Gaussian representations. As a result, PanoGaussian effectively distills panoramic exploration into Gaussian representation, preserving both global panorama exploration and the physical priors of Gaussian.

5 Experiments

5.1 Experimental Settings

Datasets. We adopt two standard benchmarks: the DyCheck IPhone dataset [39] and the Nvidia Dynamic dataset [87]. DyCheck contains long monocular video sequences with significant camera motion and challenging unseen regions in the testing split (i.e., test camera trajectories deviate significantly from the training views), allowing us to assess the exploration capabilities of PanoGaussian. We further demonstrate the generalization of PanoGaussian on in-the-wild video sequences to verify robustness under unconstrained settings.

Metrics. We evaluate reconstruction quality using standard metrics: PSNR, SSIM, and LPIPS. Following DyCheck, we also report metrics (mPSNR, mSSIM,

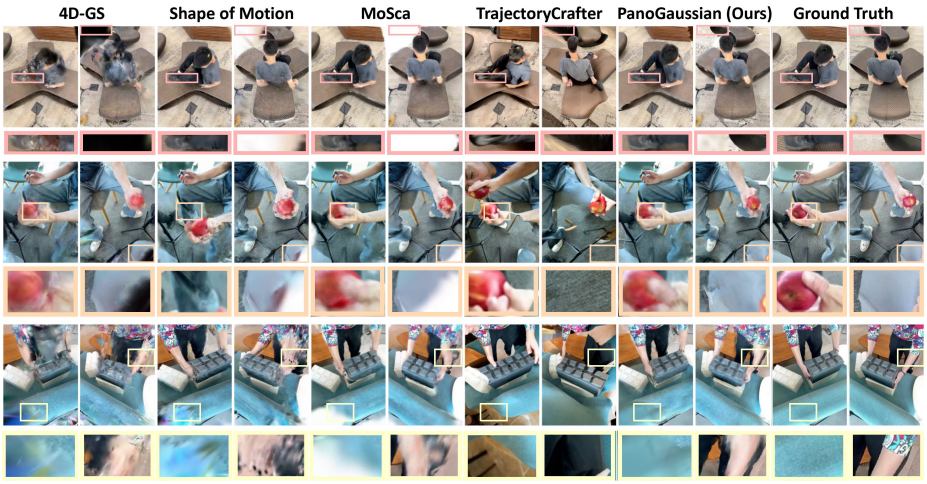

Fig. 3: Qualitative comparison of PanoGaussian on DyCheck [39]. The red, orange, and yellow boxes highlight detailed synthesis regions in each case. Previous state-of-the-art reconstruction-based approaches such as 4D Gaussian Splatting [79], Shape-of-Motion [75] and MoSca [32] struggle to infer unseen regions. These methods lead to missing (white) areas and blurred reconstructions in boxed areas. Although the generative method TrajectoryCrafter [90] can infer unseen regions, it often produces geometrically inconsistent results. In comparison, PanoGaussian achieves both geometry-consistent and complete 4D scene synthesis.

mLPIPS) on covisible regions. Since our task extends view interpolation to full 4D scene synthesis with unseen regions, we introduce additional metrics (uPSNR, uSSIM, uLPIPS) to measure performance exclusively on purely unseen areas. These unseen metrics are computed on the inverse of the DyCheck masks, specifically isolating regions outside the training camera covisibility.

Baselines. We compare our method with a broad range of monocular 4D reconstruction approaches. NeRF-based methods [13, 16, 38, 51, 54–57, 68, 69] extend NeRFs to model temporal dynamics. Gaussian Splatting methods [19, 32, 61, 75, 79, 96] achieve efficient spatio-temporal reconstruction through dynamic Gaussian primitives. Diffusion-based approaches [6, 80, 90] leverage camera-controlled video diffusion models for novel-view generation. We also compare to feed-forward methods [40, 42, 83] that directly estimate Gaussian Splatting attributes without per-scene optimization. Many of these approaches are concurrent with our work [6, 19, 40, 42, 61, 83], exist only as preprints, or have not made their code publicly available. We report on their results as provided in their respective papers. Note that recent camera-conditioned models (e.g., Gen3C [59], Recammaster [2]) are not designed for reconstruction, thus are excluded. In our tables, '-' indicates unavailable data due to unreleased full code.

Table 1: Comparison with reconstruction-based and generation-based methods on the DyCheck iPhone dataset [39]. "G.S.", "Diff", "Feed" means Gaussian Splatting methods, diffusion-based approaches, feed-forward methods, respectively. Best, second-best, and third-best results are highlighted in red, orange, and yellow, respectively. Our model consistently achieves the best performance in all metrics.

Method	Type	PSNR ↑	SSIM ↓	LPIPS ↑	mPSNR ↑	mSSIM ↓	mLPIPS ↑	uPSNR ↑	uSSIM ↓	uLPIPS
4D G.S. [79]	G.S.	15.71	0.450	0.398	16.54	0.594	0.347	12.82	0.302	0.452
Shape-of-Motion [75]	G.S.	16.79	0.510	0.391	17.32	0.598	0.296	14.84	0.321	0.448
MoSca [32]	G.S.	17.31	0.573	0.354	19.32	0.706	0.264	14.23	0.295	0.461
BulletGen [61]	G.S.	–	–	–	16.78	0.640	0.390	–	–	–
Vivid4D [28]	G.S.	–	–	–	15.20	0.500	0.493	–	–	–
Cat4D [80]	Diff.	15.91	0.427	0.398	17.39	0.607	0.341	14.06	0.351	0.428
TrajectoryCrafter [90]	Diff.	13.58	0.373	0.483	15.48	0.523	0.426	10.88	0.315	0.441
CogNVS [6]	Diff.	16.94	0.449	0.598	18.63	0.614	0.290	14.97	0.371	0.487
MoVieS [42]	Feed	–	–	–	18.46	0.589	0.309	–	–	–
BTimer [40]	Feed	–	–	–	16.52	0.570	0.338	–	–	–
4DGT [83]	Feed	15.04	0.446	0.423	16.12	0.556	0.408	12.11	0.301	0.450
4D-Fly [77]	Feed	–	–	–	17.03	0.60	0.37	–	–	–
PanoGaussian (Ours)	Pano	**18.71**	**0.598**	**0.323**	**19.85**	**0.741**	**0.239**	**16.72**	**0.495**	**0.385**

Implementation details. We implement our panoramic trajectory training based on the pretrained TrajectoryCrafter [90]. Training is conducted on the large-scale OpenVid dataset [52], which provides diverse dynamic scenes and high-fidelity visual details suitable for optimizing generative priors. The video frames are preprocessed to a resolution of 384×672 and a length of 49 frames. To avoid overfitting, only the cross-attention and patch embedding layers in the Ref-DiT blocks are optimized. The whole training takes 25,000 iterations with a learning rate of 2×10^{-6}. For Panoramic Paradigm Inference, we adopt a spherical coordinate system with azimuthal angle $\phi \in [-\pi, \pi)$ and polar angle $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$. The panorama resolution is set to $w_p = 2048$ and $h_p = 2048$. We use $M = 8$ directional views for full-scene coverage, $E = 6$ expansion steps, and an angular step of $\alpha = 15°$ in the progressive expansion strategy. We employ MoSca [32] as the dynamic Gaussian backbone for our method and all Gaussian-based comparisons to ensure fair evaluation. The refinement weight is set to $\lambda_{\text{refine}} = 0.1$. Gaussian optimization runs for 10,000 iterations. Specifically, every 2,000 iterations of GS optimization, we perform progressive panoramic expansion via the video diffusion prior. Total per-scene optimization time is approximately 2.0h, rendering at 32.42FPS. Our 2.0h runtime consists of 1.4h for 4D Gaussian Splatting fitting and only 0.6h for progressive diffusion generation. The generation time scales linearly with the number of expansion steps E and tangent directions M. Compared to baselines, this is significantly faster than Shape of Motion [75] (4.2h) and comparable to MoSca [32] (1.3h). For evaluation on DyCheck and Nvidia datasets, we report quantitative results using Gaussian representation renderings. For extreme view exploration, we present the refined final-step results in the panoramic representation to emphasize visual quality.

Table 2: Quantitative comparison with reconstruction-based methods on Nvidia [39]. "NeRF", "G.S.", "Feed" means NeRF-based, Gaussian Splatting, and feed-forward methods, respectively. Best, second-best, and third-best results are highlighted in red, orange, and yellow, respectively.

Method	Type	PSNR↑	LPIPS↓	Method	Type	PSNR↑	LPIPS↓
NSFF [38]	NeRF	24.33	0.199	4D G.S. [79]	G.S.	21.45	0.199
D-NeRF [56]	NeRF	21.49	0.232	GaussianMarbles [66]	G.S.	22.32	0.129
DynNeRF [16]	NeRF	26.10	0.082	MoSca [32]	G.S.	26.72	0.070
MonoNeRF [68]	NeRF	25.62	0.106	BulletGen [61]	G.S.	17.02	0.386
CTNeRF [51]	NeRF	26.13	0.082	MoVieS [42]	Feed	19.16	0.315
HyperNeRF [55]	NeRF	17.60	0.367	4D-Fly [77]	Feed	22.52	0.14
DynPoint [96]	NeRF	26.53	**0.068**	PanoGaussian (Ours)	Pano	**26.75**	0.069

5.2 Novel View Synthesis

Results on DyCheck [39]. As shown in Tab. 1 and Fig. 3, PanoGaussian consistently achieves the best performance in all metrics. NeRF-based and Gaussian Splatting methods primarily focus on view interpolation within the observed camera trajectories. As a result, they cannot infer unseen regions, leading to missing reconstructions in the boxed areas of Fig. 3. This limitation also causes a notable performance drop when switching from mPSNR (covisible regions) to PSNR (full scene) in Tab. 1. Video diffusion approaches, on the other hand, often produce geometrically distorted results due to the absence of explicit 3D priors. In comparison, PanoGaussian leverages a unified Panoramic-Gaussian representation to achieve geometry-consistent 4D scene synthesis, demonstrating superior structural fidelity even under viewpoint changes.

Results on Nvidia [87]. Unlike DyCheck, the Nvidia testing split consists of scenes with minimal viewpoint variation and no truly unseen regions. Consequently, PanoGaussian achieves performance comparable to existing methods, as shown in Fig. 5 and Tab. 2. The extrapolation gains observed on DyCheck do not fully manifest here. This result is expected, since our approach is primarily designed to explore and reconstruct unseen regions, while also preserving interpolation quality within observed viewpoints. To better highlight this strength, we additionally perform an extreme view exploration experiment on the Nvidia dataset without ground-truth supervision, where PanoGaussian qualitatively synthesizes coherent and geometrically consistent novel views.

Qualitative Results on Kubric-4D [71]. Unlike the Nvidia dataset, Kubric-4D features substantial camera motion with genuinely unseen regions, making it a more suitable benchmark for evaluating extrapolation. As shown in Fig. 4, reconstruction-based and diffusion-based baselines produce incoherent or missing content in unseen regions, whereas PanoGaussian maintains structurally and visually consistent synthesis.

Fig. 4: Qualitative comparison on Kubric-4D [71]. Under large camera displacement, baselines fail to synthesize coherent unseen regions, while PanoGaussian preserves consistent structure and appearance.

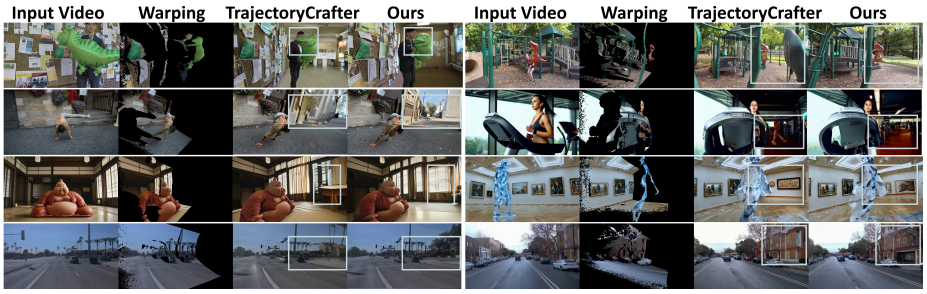

Fig. 5: Scene exploration under extreme camera transitions. We visualize warping results under camera trajectories exhibiting drastic viewpoint changes beyond the original range on Nvidia [87] and in-the-wild dynamic videos. Previous method [90] produces severe geometric distortions and artifacts in the green boxed areas. In comparison, PanoGaussian preserves geometry and synthesizes consistent and realistic scenes even under extreme camera transitions.

Extreme View Exploration. To evaluate the generalizability of PanoGaussian, we test its robustness under extreme camera transitions far beyond the original trajectory. We conduct scene exploration on in-the-wild dynamic videos containing multiple objects and complex motions. Since extreme in-the-wild scene exploration lacks ground truth, quantitative evaluation is infeasible. Furthermore, we focus our qualitative comparison on the most relevant state-of-the-art TrajectoryCrafter [90]. NeRF-based and Gaussian-based reconstruction methods fail to infer unseen regions in this challenging situation without supervision. As illustrated in Fig. 5, PanoGaussian achieves consistent 4D scene synthesis under a drastic viewpoint transition. In comparison, previous video diffusion methods [90] suffer from severe geometric distortion, such as the deformed balloon in the first row, physically inconsistent buildings in the second and third rows, and the distorted human in the last row. These results demonstrate that our Panoramic-Gaussian framework effectively unifies the strengths of geometric reconstruction and video generative priors for robust 4D scene synthesis, enabling physically consistent 4D scene synthesis.

Fig. 6: Panoramic design is vital for unseen-region consistency. Planar expansion with the same 4DGS backbone fails in unseen regions, while panoramic coordinates recover coherent geometry even without full Gaussian refinement.

5.3 Ablation Studies

Necessity of Panoramic Representation. To validate the necessity of our panoramic design, we disentangle panoramic expansion from 4DGS refinement on DyCheck [39]. We compare planar expansion with the same 4DGS backbone, panoramic expansion without Gaussian refinement, and our full model in Tab. 3 and Fig. 6. As shown in Tab. 3, planar expansion with the identical backbone substantially underperforms our full model on uPSNR, confirming that panoramic design is the prerequisite for unseen-region reconstruction. Panoramic expansion alone already recovers most of this gap, while 4DGS refinement further improves overall visual quality in both seen and unseen regions. As shown in Fig. 6, the planar baseline fails to capture holistic scene structure, leading to severe geometric distortion and multi-view inconsistency when resolving large occlusions. Our panoramic representation ensures physically coherent unconstrained 4D scene exploration.

Panoramic Trajectories Training. Training in panoramic trajectories provides limited quantitative gains in DyCheck [39] in Tab. 3. However, it yields clear qualitative improvements under significant viewpoint changes. As shown in Fig. 7, optimization along the designed panoramic trajectories produces results more consistent with the original scene. This is mainly attributed to the consistency in training and inference of PanoGaussian, which enables a more stable and coherent view exploration.

Progressive Expansion. Our proposed progressive expansion strategy has significant contributions to explore and infer consistent scenes. In Fig. 7, without this progressive strategy, direct one-step generation often relies on limited visual contexts, leading to fragmented or inconsistent results. In contrast, our progressive expansion gradually enlarges the explored region, allowing the model to reason over more comprehensive information. This process effectively constrains generative priors and leads to smoother and more consistent reconstructions, especially under large camera transitions. This strategy has also made a significant contribution to qualitative results in Tab. 3.

Table 3: Quantitative ablation on different components of the system. Best, second-best, and third-best results are highlighted in red, orange, and yellow, respectively. Our proposed Panoramic representation and Gaussian Splatting representation yields the most significant quantitative improvement. The progressive expansion strategy further contributes to stable and coherent reconstruction.

Method	↑ PSNR	↑ SSIM	↓ LPIPS	↑ uPSNR	↑ uSSIM	↓ uLPIPS
Planar with 4DGS	16.47	0.550	0.377	14.25	0.302	0.445
Pano without 4DGS refinement	17.12	0.555	0.358	15.08	0.421	0.418
Without Panoramic Training	18.46	0.584	0.327	16.32	0.476	0.397
Without Progressive Expansion	17.51	0.576	0.351	15.22	0.446	0.417
Without Gaussian Splatting	14.67	0.392	0.445	12.61	0.314	0.482
PanoGaussian (Ours)	**18.71**	**0.598**	**0.323**	**16.72**	**0.495**	**0.385**

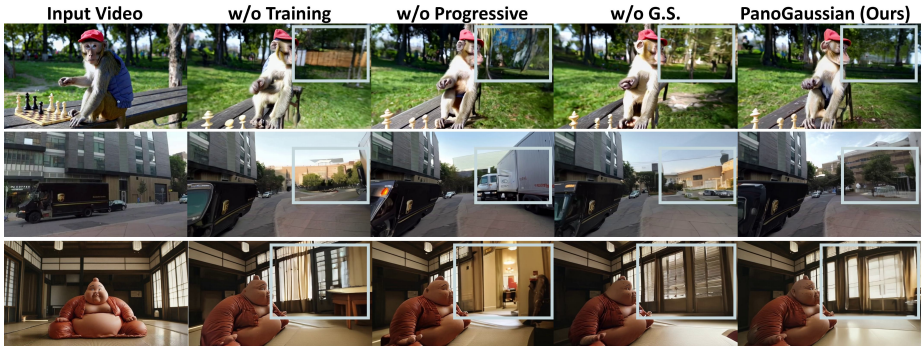

Fig. 7: Qualitative ablation study under extreme camera transitions. The green boxed areas highlight regions with notable differences. Training with panoramic trajectories and a progressive expansion strategy enables PanoGaussian to infer unseen regions more consistently. The unification with Gaussian Splatting representation further preserves physical realism and geometric fidelity.

Gaussian Splatting Representation. As shown in Tab. 3, the injection of the Gaussian Splatting representation significantly contributes to the improvement of qualitative results in the DyCheck IPhone dataset. The geometric priors from Dynamic Gaussian Splatting encourage faithful alignment with the original monocular video. This geometric alignment significantly reduces structural distortions and unrealistic artifacts, which improves physical consistency with the original monocular videos. As shown in Fig. 7, a visual failure example without Gaussian Splatting clearly exhibits temporal inconsistency, while our full unification improves both visual fidelity and geometric integrity, offering the physical consistency of 4D scenes.

Error Propagation Analysis. Our design structurally prevents cascading errors during progressive expansion. At every expansion step, warping is performed

Table 4: Reconstruction quality across expansion steps on DyCheck [39]. It confirms that errors do not accumulate during progressive expansion.

Expansion Step	6	12	18	24
PSNR / uPSNR (dB)	18.71 / 16.72	18.72 / 16.74	18.69 / 16.68	18.68 / 16.66

from the *original* observed video $V_{\text{src,inf}}$ using fixed 4D GS depth and poses, rather than from previously hallucinated outputs. The binary mask M_e restricts both diffusion inpainting and refinement loss to newly generated regions only, preventing re-optimization of already settled areas. As shown in Tab. 4, PSNR and uPSNR remain stable across expansion steps, confirming that each step's quality is independently determined by the fixed 4D GS reconstruction rather than accumulated from prior hallucinations.

6 Conclusion

We propose PanoGaussian, a unified Panoramic-Gaussian representation for the synthesis of geometry-aware 4D scenes. The panoramic representation, equipped with a progressive expansion strategy, enables a comprehensive exploration of the whole scene. The Gaussian representation ensures high geometric fidelity by explicitly modeling spatial structures. Extensive experiments demonstrate that PanoGaussian achieves consistent 4D scene synthesis, maintaining strong geometric accuracy even under large viewpoint variations. We expect that this unified representation may facilitate future work on physically grounded and controllable 4D content generation.

Limitations. PanoGaussian is most effective on inward-facing, object-centric scenes with sufficient inter-view overlap to anchor geometry. On outward-facing captures, distant content provides weaker multi-view constraints and quality becomes less reliable. The pipeline further relies on a pretrained video diffusion model and per-scene 4DGS optimization ($\sim$2.0 h per scene), which limits real-time deployment and may inherit artifacts from the generative prior.

Acknowledgement.

This work was supported in part by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123004), Ningbo grant (2025Z038) and National Natural Science Foundation of China (Grant No. 62376060).

References

1. Ai, H., Cao, Z., Lu, H., Chen, C., Ma, J., Zhou, P., Kim, T.K., Hui, P., Wang, L.: Dream360: Diverse and immersive outdoor virtual scene creation via transformer-based 360 image outpainting. IEEE transactions on visualization and computer graphics (2024)
2. 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 (2025)
3. Bai, J., Huang, L., Guo, J., Gong, W., Li, Y., Guo, Y.: 360-gs: Layout-guided panoramic gaussian splatting for indoor roaming. In: 3DV (2025)
4. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
5. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In: SIGGRAPH (2025)
6. Chen, K., Khurana, T., Ramanan, D.: Reconstruct, inpaint, finetune: Dynamic novel-view synthesis from monocular videos. In: NeurIPS (2025)
7. Chen, Z., Wu, C., Shen, Z., Zhao, C., Ye, W., Feng, H., Ding, E., Zhang, S.H.: Splatter-360: Generalizable 360 gaussian splatting for wide-baseline panoramic images. In: CVPR (2025)
8. Chu, W.H., Ke, L., Fragkiadaki, K.: Dreamscene4d: Dynamic multi-object scene generation from monocular videos. In: NeurIPS (2024)
9. Chugunov, I., Joshi, A., Murthy, K., Bleibel, F., Heide, F.: Neural light spheres for implicit image stitching and view synthesis. In: SIGGRAPH (2024)
10. Das, D., Wewer, C., Yunus, R., Ilg, E., Lenssen, J.E.: Neural parametric gaussians for monocular non-rigid object reconstruction. arXiv preprint arXiv:2312.01196 (2023)
11. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. ArXiv (2024)
12. Duisterhof, B., Xu, D., Chen, R., Chen, Y.C.: Mvsplat: Efficient 3d gaussian splatting from multi-view stereo. arXiv preprint arXiv:2312.01089 (2023)
13. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH (2022)
14. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: CVPR (2023)
15. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
16. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: ICCV (2021)
17. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Monocular dynamic view synthesis: A reality check. In: NeurIPS (2022)
18. Gu, K., Maugey, T., Knorr, S., Guillemot, C.: Omni-nerf: neural radiance field from 360 image captures. In: ICME (2022)
19. Guo, F., Hsu, C.C., Ding, S., Zhang, C.: Uncertainty matters in dynamic gaussian splatting for monocular 4d reconstruction. ICLR (2026)

20. Guo, Z., Zhou, W.g., Li, L., Wang, M., Li, H.: Motion-aware 3d gaussian splatting for efficient dynamic scene reconstruction. ArXiv (2024)
21. He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., Yang, C.: Cameractrl: Enabling camera control for text-to-video generation. arXiv preprint (2024)
22. He, H., Yang, C., Lin, S., Xu, Y., Wei, M., Gui, L., Zhao, Q., Wetzstein, G., Jiang, L., Li, H.: Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. arXiv preprint (2025)
23. Hedman, P., Alsisan, S., Szeliski, R., Kopf, J.: Casual 3d photography. ACM TOG (2017)
24. Hedman, P., Kopf, J.: Instant 3d photography. ACM TOG (2018)
25. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
26. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers. ICLR (2022)
27. Huang, H., Chen, Y., Zhang, T., Yeung, S.K.: 360roam: Real-time indoor roaming using geometry-aware 360 radiance fields. arXiv preprint arXiv:2208.02705 (2022)
28. Huang, J., Miao, S., Yang, B., Ma, Y., Liao, Y.: Vivid4d: Improving 4d reconstruction from monocular video by video inpainting. In: ICCV (2025)
29. Katsumata, K., Vo, D.M., Nakayama, H.: An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. ArXiv (2023)
30. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (2023)
31. Lee, S., Chung, J., Huh, J., Lee, K.M.: Odgs: 3d scene reconstruction from omnidirectional images with 3d gaussian splattings. NeurIPS (2024)
32. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: CVPR (2025)
33. Li, F., Zhang, H., Ahuja, N.: Self-calibrating 4d novel view synthesis from monocular videos using gaussian splatting. ArXiv (2024)
34. Li, L., Huang, H., Yeung, S.K., Cheng, H.: Omnigs: Fast radiance field reconstruction using omnidirectional gaussian splatting. In: WACV (2025)
35. Li, R., Pan, P., Yang, B., Xu, D., Zhou, S., Zhang, X., Li, Z., Kadambi, A., Wang, Z., Tu, Z., et al.: 4k4dgen: Panoramic 4d generation at 4k resolution. arXiv preprint arXiv:2406.13527 (2024)
36. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. arXiv preprint arXiv:2312.16812 (2023)
37. Li, Z., Niklaus, S., Snavely, N., Wang, O.: 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 (2021)
38. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: CVPR (2021)
39. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: Neural dynamic image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
40. Liang, H., Ren, J., Mirzaei, A., Torralba, A., Liu, Z., Gilitschenski, I., Fidler, S., Oztireli, C., Ling, H., Gojcic, Z., et al.: Feed-forward bullet-time reconstruction of dynamic scenes from monocular videos. arXiv preprint (2024)
41. Liang, Y., Khan, N., Li, Z., Nguyen-Phuoc, T., Lanman, D., Tompkin, J., Xiao, L.: Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. ArXiv (2023)

42. 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)
43. 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 preprint arXiv:2506.09997 (2025)
44. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. arXiv:2312.03431 (2023)
45. Liu, J., Lin, S., Li, Y., Yang, M.H.: Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In: CVPR (2025)
46. Lu, Z., Zheng, Q., Shi, B., Jiang, X.: Pano-nerf: Synthesizing high dynamic range novel views with geometry from sparse low dynamic range panoramic images. In: AAAI (2024)
47. Lu, Z., Hu, K., Wang, C., Bai, L., Wang, Z.: Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In: AAAI (2024)
48. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 3DV (2024)
49. Luo, Y., Bai, J., Shi, X., Xia, M., Wang, X., Wan, P., Zhang, D., Gai, K., Xue, T.: Camclonemaster: Enabling reference-based camera control for video generation. arXiv preprint arXiv:2506.03140 (2025)
50. Luo, Z., Ran, H., Lu, L.: Instant4d: 4d gaussian splatting in minutes. arXiv preprint arXiv:2510.01119 (2025)
51. Miao, X., Bai, Y., Duan, H., Wan, F., Huang, Y., Long, Y., Zheng, Y.: Ctnerf: Cross-time transformer for dynamic neural radiance field from monocular video. PR (2024)
52. Nan, K., Xie, R., Zhou, P., Fan, T., Yang, Z., Chen, Z., Li, X., Yang, J., Tai, Y.: Openvid-1m: A large-scale high-quality dataset for text-to-video generation. arXiv preprint arXiv:2407.02371 (2024)
53. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: ICML (2021)
54. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: ICCV (2021)
55. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: a higher-dimensional representation for topologically varying neural radiance fields. ACM TOG (2021)
56. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
57. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: CVPR (2021)
58. Ramasinghe, S., Shevchenko, V., Avraham, G., van den Hengel, A.: Blirf: Band limited radiance fields for dynamic scene modeling. In: AAAI (2024)
59. 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: CVPR (2025)
60. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
61. Rozumnyi, D., Luiten, J., Khan, N., Schonberger, J., Kontschieder, P.: Bullet-gen: Improving 4d reconstruction with bullet-time generation. arXiv preprint arXiv:2506.18601 (2025)

62. Shade, J., Gortler, S., He, L.w., Szeliski, R.: Layered depth images. In: Proceedings of the 25th annual conference on Computer graphics and interactive techniques (1998)
63. Shen, Q., Yi, X., Lin, M., Zhang, H., Yan, S., Wang, X.: Seeing world dynamics in a nutshell. arXiv preprint arXiv:2502.03465 (2025)
64. Shum, H.Y., He, L.W.: Rendering with concentric mosaics. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques (1999)
65. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. IEEE Transactions on Visualization and Computer Graphics (2023)
66. Stearns, C., Harley, A., Uy, M., Dubost, F., Tombari, F., Wetzstein, G., Guibas, L.: Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In: SIGGRAPH (2024)
67. Sun, J., Jiao, H., Li, G., Zhang, Z., Zhao, L., Xing, W.: 3dgstream: On-the-fly training of 3d gaussians for efficient streaming of photo-realistic free-viewpoint videos. arXiv preprint arXiv:2403.01444 (2024)
68. Tian, F., Du, S., Duan, Y.: Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In: ICCV (2023)
69. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: ICCV (2021)
70. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: ECCV (2024)
71. Van Hoorick, B., Wu, R., Ozguroglu, E., Sargent, K., Liu, R., Tokmakov, P., Dave, A., Zheng, C., Vondrick, C.: Generative camera dolly: Extreme monocular dynamic novel view synthesis. In: ECCV (2024)
72. Wang, C., Eckart, B., Lucey, S., Gallo, O.: Neural trajectory fields for dynamic novel view synthesis. In: ArXiv Preprint (2021)
73. Wang, G., Wang, P., Chen, Z., Wang, W., Loy, C.C., Liu, Z.: Perf: Panoramic neural radiance field from a single panorama. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
74. Wang, Q., Li, W., Mou, C., Cheng, X., Zhang, J.: 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In: CVPR (2024)
75. 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 (2025)
76. Wang, S., Yang, X., Shen, Q., Jiang, Z., Wang, X.: Gflow: Recovering 4d world from monocular video. In: AAAI (2025)
77. 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 (2025)
78. 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. arXiv preprint arXiv:2310.08528 (2023)
79. 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 (2024)
80. 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. arXiv preprint (2024)
81. Wu, T., Zheng, C., Cham, T.J.: Panodiffusion: 360-degree panorama outpainting via diffusion. arXiv preprint arXiv:2307.03177 (2023)

82. Wu, T., Yang, S., Po, R., Xu, Y., Liu, Z., Lin, D., Wetzstein, G.: Video world models with long-term spatial memory. arXiv preprint (2025)
83. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. arXiv preprint arXiv:2506.08015 (2025)
84. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In: International Conference on Learning Representations (ICLR) (2024)
85. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023)
86. Yesiltepe, H., Yanardag, P.: Dynamic view synthesis as an inverse problem. arXiv preprint (2025)
87. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: CVPR (2020)
88. You, M., Zhu, Z., Liu, H., Hou, J.: Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. In: ICLR (2025)
89. Yu, H., Julin, J., Milacski, Z.A., Niinuma, K., Jeni, L.A.: Cogs: Controllable gaussian splatting. arXiv (2023)
90. YU, M., Hu, W., Xing, J., Shan, Y.: Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. arXiv preprint (2025)
91. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023)
92. Zhang, Q., Song, J., Huang, X., Chen, Y., Liu, M.Y.: Diffcollage: Parallel generation of large content with diffusion models. In: CVPR (2023)
93. Zheng, K.C., Kang, S.B., Cohen, M.F., Szeliski, R.: Layered depth panoramas. In: CVPR (2007)
94. Zheng, S., Yin, M., Hu, W., Li, X., Shan, Y., Fu, Y.: Versecrafter: Dynamic realistic video world model with 4d geometric control. In: CVPR (2026)
95. Zhou, J.J., Gao, H., Voleti, V., Vasishta, A., Yao, C.H., Boss, M., Torr, P., Rupprecht, C., Jampani, V.: Stable virtual camera: Generative view synthesis with diffusion models. arXiv preprint (2025)
96. Zhou, K., Zhong, J.X., Shin, S., Lu, K., Yang, Y., Markham, A., Trigoni, N.: Dynpoint: Dynamic neural point for view synthesis. In: NeurIPS (2024)
97. Zhou, S., Ren, H., Weng, Y., Zhang, S., Wang, Z., Xu, D., Fan, Z., You, S., Wang, Z., Guibas, L., et al.: Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature fields. In: CVPR (2025)