This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Robust Multi-Object 4D Generation for In-the-wild Videos

Wen-Hsuan Chu[1*] Lei Ke[1*] Jianmeng Liu[1*] Mingxiao Huo[1]
Pavel Tokmakov[2] Katerina Fragkiadaki[1]
[1]Carnegie Mellon University [2]Toyota Research Institute

Abstract

We address the challenge of generating dynamic 4D scenes from monocular multi-object videos with heavy occlusions and introduce GenMOJO, a novel approach that integrates rendering-based deformable 3D Gaussian optimization with generative priors for view synthesis. While existing view-synthesis models excel at novel view generation for isolated objects, they struggle with full scenes due to their complexity and data demands. To overcome this, GenMOJO decomposes scenes into individual objects, optimizing a differentiable set of deformable Gaussians per object while capturing 2D occlusions from a 3D perspective through joint Gaussian splatting. Joint splatting ensures occlusion-aware rendering losses in observed frames, while explicit object decomposition allows the usage of object-centric diffusion models for object completion in unobserved viewpoints. To reconcile the differences between object-centric priors and the global frame-centric coordinate system of the video, GenMOJO employs differentiable transformations to unify the rendering and generative constraints within a single framework. The result is a model capable of generating 4D objects across space and time while producing 2D and 3D point tracks from monocular videos. To rigorously evaluate the quality of scene generation and the accuracy of the motion under multi-object occlusions, we introduce MOSE-PTS, a subset of the challenging MOSE benchmark, which we annotated with high-quality 2D point tracks. Quantitative evaluations and perceptual human studies confirm that GenMOJO generates more realistic novel views of scenes and produces more accurate point tracks compared to existing approaches. Project page: https://genmojo.github.io/.

1. Introduction

Humans effortlessly perceive complete, coherent objects and their motions from videos even though they only observe the pixels on the front surface of a dynamic scene, which constantly changes due to motion, occlusion, and reappearance. Recovering persistent and accurate 4D representations from monocular videos with multiple objects is a highly under-constrained problem. However, it is critical for fine-grained video understanding, visual imitation in robotics [12], and building causal world models of intuitive physics [50] that can simulate the consequences of actions.

Recent view-predictive generative models [24, 40] provide powerful priors for novel view synthesis. However, existing video-to-4D generation works [5, 14, 55, 56] based on these generative models often struggle to maintain temporal coherence and accurate 3D geometry in complex multi-object scenes due to the intricate dynamics involved, such as heavy occlusions and fast motion.

In this paper, we propose GenMOJO, a compositional method that enables **M**ulti-**O**bject **JO**int splatting for 4D scene **GEN**eration from monocular videos by combining frame-centric test-time rendering error optimization with object-centric generative priors. GenMOJO decomposes a scene into individual foreground objects and background, representing each object with a differentiable and deformable set of 3D Gaussians. We optimize the parameters of these Gaussians by: (1) jointly splatting all Gaussians to compute rendering errors in the input video frames, accounting for occlusions; and (2) individually rendering each object's Gaussians from unobserved viewpoints to optimize a score distillation (SDS) objective. To bridge frame-centric and object-centric coordinate frames, GenMOJO employs differentiable affine transformations that align the object-centric Gaussians with the frame-centric Gaussians.

The work most closely related to ours is DreamScene4D [5], which also leverages object-centric generative priors to produce 4D scene completions from monocular video. However, their approach inpaints and optimizes the 3D deformable Gaussians for each object independently and only combine them during a final rendering step to infer object depth ordering. This strategy struggles with cross-object occlusions, as rendering objects separately fails to account for occluders. As a result, their method often produces unnatural, inter-penetrating artifacts in scenes with multiple heavily occluded objects, as demonstrated in our experiments. In contrast, our approach unifies

*Equal contributions



Figure 1. **GenMOJO** extends video-to-4D generation to complex, real-world scenarios characterized by heavy multi-object occlusions and rapid motion, enabling the creation of complete 4D scene representations with 360° novel view synthesis and accurate point motion tracking. We showcase rendered images from diverse viewpoints and motion trajectories (visualizing a single object for clarity) in the re-rendered images across different time steps.

image-based and object-centric constraints within a single gradient-based optimization framework. By jointly splatting all object Gaussians, we accurately model cross-object occlusions and interactions, while still imposing strong appearance priors by rendering each object independently.

As shown in Figure 1, GenMOJO emphasizes addressing real-world complexity through the evaluation of challenging datasets with frequent object interactions. Unlike previous methods that focused on single-object [14, 37] or simple multi-object videos [5], GenMOJO is designed to handle scenes with heavy occlusions, crowded environments, and intricate dynamics.

We evaluate GenMOJO on multi-object videos from DAVIS and a subset of MOSE [6], for which we manually annotated accurate point tracks. These videos feature complex multi-object interactions and frequent real-world occlusions. We compare against state-of-the-art NeRF- and Gaussian-based 4D generation methods [5, 14, 47]. Our results show substantial improvements in both 4D rendering quality and motion accuracy. In addition, we conducted human evaluations to assess the perceptual realism of our results, which further support our quantitative findings.

2. Related work

Dynamic Scene Reconstruction. Dynamic scene reconstruction aims to lift the visible parts of a video to a dynamic 3D representation for novel view synthesis. These methods often take videos where a large number of camera views exist as input [27, 49, 54], alleviating the issue of partial observability. As a result, many of these methods can not be readily applied to monocular videos, as no constraints exist in unobserved viewpoints. These methods fall into two large categories. Dynamic Neural Radiance Field (NeRF) based methods [3, 21, 25, 31, 34] extend NeRFs [29] to dynamic scenes and model a dynamic scene using grid-based, voxel-based representations [4, 23, 26] or through deformation networks [4, 8]. Recently, Gaussian Splatting [18, 27, 49] has gained great attention in the field of neural rendering. Compared to NeRF-based methods, Gaussian Splatting represents the scene as explicit Gaussians and uses efficient rasterization techniques for rendering, greatly improving the computation efficiency. There have also been efforts to model the visible parts of a video scene from monocular inputs, but they do not attempt to model the unobserved parts [20, 44, 47, 48]. Compared to these works, we adopt dynamic Gaussians to synthesize both the visible and unobserved parts of the video scene.

Dynamic Scene Generation. In contrast to dynamic scene reconstruction, dynamic scene generation is concerned with modeling a complete video scene across both visible and unobserved viewpoints, often operating under monocular or sparse view settings. Existing text-to-4D [1, 2, 22, 43] or image-to-4D [37, 53, 59] works use dif-

fusion models [24] to condition on text or image inputs and generate synthetic videos, then use score distillation sampling (SDS) [33] losses to supply constraints in unobserved viewpoints to synthesize complete 4D representations using dynamic NeRFs or Gaussians. Many existing video-to-4D works simplify this problem by assuming that the input consists of a single object undergoing small movements [9, 14, 30, 37, 55, 56]. More recently, this line of work has been extended to multi-object video by first decomposing the video into multiple object tracks, optimizing them independently, and then using depth models [51] to recompose the optimized 4D objects to re-form the original video scene [5]. However, the late multi-object composition strategy in [5] is extremely reliant on the quality of the depth model, and the optimization process itself is unaware of the occlusions that occur in the observed video sequence. In comparison, we propose to compose early and optimize the deformations of the objects *jointly*, which naturally allows the optimization process to reason about the cross-object occlusions that occur in videos.

Point Motion via 4D Generation / Reconstruction. 4D generation or reconstruction from monocular videos enables the extraction of point trajectories through test-time optimization by fitting dynamic radiance fields [5, 27, 41, 46, 47] to the observed video frames without the need for point annotations. In contrast, data-driven point trackers are trained on synthetic training data, yet outperform others in existing benchmarks by a noticeable margin [15]. While our dynamic Gaussian representation is not specifically designed for precise point tracking via Gaussian Splatting, it facilitates straightforward extraction of point trajectories for motion accuracy evaluation. In particular, we observe that the gap in tracking performance between our method and feedforward approaches decreases significantly in multi-object videos with heavy occlusions [6], since the reconstruction objective bounds the tracking error.

3. Robust Multi-Object 4D Generation

To generate dynamic 4D scenes from a monocular video with complex multi-object dynamics, we introduce GenMOJO, which utilizes Gaussian Splatting [18, 49] as the 4D scene representation and leverages diffusion-based generative models [24] for novel view synthesis.

3.1. Background: Generative Gaussian Splatting

Gaussian Splatting [18] represents a scene through a set of 3D Gaussians, each characterized by its position μ, scale s, rotation q, opacity α, and spherical harmonic (SH) coefficients f to represent color. To extend 3D Gaussians for novel view synthesis, Score Distillation Sampling (SDS) [33] is commonly applied. For example, DreamGaussian [45] leverages Zero-1-to-3 [24] as the diffusion prior, which uses a reference view and a relative camera pose to generate plausible images given a target viewpoint, lifting a single frame from 2D to 3D.

To model 3D scene dynamics, existing methods [5, 37] then deform these 3D Gaussians over time. Starting with a set of 3D Gaussians derived from the initial frame, each Gaussian is parameterized at each timestep by a set of learnable variables such as its 3D position μ_t, 3D rotation (a quaternion) q_t, and 3D scale s_t. The RGB (spherical harmonics) values and opacity remain constant across timesteps, inherited from the first-frame 3D Gaussians. DreamGaussian4D [37] models object dynamics by directly optimizing Gaussian deformations in an object-centric video through a K-plane [8] based deformation network with an RGB rendering loss and a SDS loss. DreamScene4D [5] extends this approach to multi-object scenes by introducing a "decompose-recompose" strategy that optimizes 4D Gaussians for each object track independently. However, these methods struggle with multi-object occlusions since they model objects separately, without considering the interactions and occlusions between them.

3.2. GenMOJO

We introduce GenMOJO, a framework for video-to-4D generation in complex multi-object scenes, specifically designed to accurately model object interactions and occlusions. GenMOJO employs an "early-composition" strategy by first composing 3D Gaussians in a global space with depth initialization, then optimizing for their temporal deformations. Furthermore, we render instance masks to preserve Gaussian instance identities during the joint splatting process, effectively managing significant 2D mutual occlusions between objects by leveraging the compositional structure of the 3D scene.

As illustrated in Figure 2, our method processes a monocular multi-object video by first isolating each object using state-of-the-art segmenters and trackers [17, 36, 38]. We then apply Score Distillation Sampling (SDS) [33] with diffusion priors to generate static 3D Gaussian representations for each object. Unlike DreamScene4D [5], which independently optimizes Gaussian deformations and later composes the objects, GenMOJO composes the entire 3D scene upfront using initial depth predictions. This enables a joint optimization of all Gaussian deformations, enhancing stability and overall generation quality.

3.2.1. Multi-object Joint Splatting

After independently optimizing 3D Gaussians for each object in the initial frame, we compose them within a unified coordinate frame to create a coherent 3D scene with initial depth ordering. This process involves determining the depth and scaling of each object along camera rays. Specifically, we employ a pre-trained consistent video depth estimator [13] to determine the relative depth of each object,

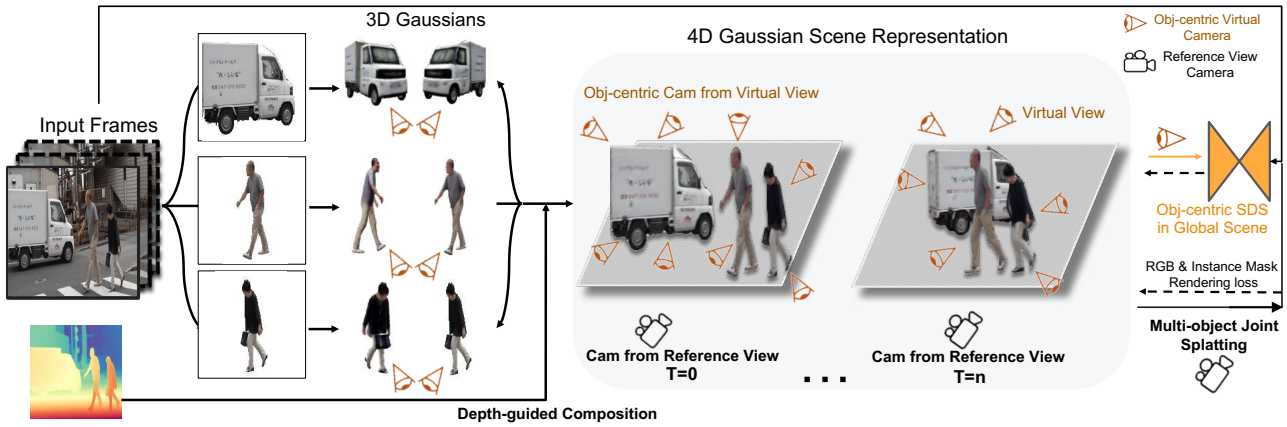


Figure 2. **Method overview of GenMOJO.** Starting with a monocular video, GenMOJO uses segmentation and tracking methods [36, 38] to isolate each object. Score Distillation Sampling (SDS) [33] is applied to generate static 3D Gaussian representations. The 3D Gaussians for each object are first composed in global space with depth-based initialization [13] and optimized using re-projection error from the video, along with individual object renders for score distillation. Differentiable affine transformations are employed to jointly optimize both image-centric reprojection and object-centric score distillation objectives within a unified framework.

which then guides the composition of the scene. We randomly select a "reference" object j and compute the relative depth scaling for every other object i as $k_i = \frac{D_i}{D_j}$ in each frame, where D_i and D_j is the median depth for objects i and j.

In the t-th frame, the original 3D position μ_t^i and the scale $\mathbf{s}_t^i$ of the Gaussians for object i are scaled along the camera rays using the scaling factor k_i: $\mathbf{p}_t^i = \mathcal{C}^r - \left(\mathcal{C}^r - \mu_t^i\right) \times k_i$ and $\mathbf{s}_t^i = \mathbf{s}_t^i \times k_i$, where $\mathbf{p}$ is the updated 3D position and $\mathcal{C}^r$ denotes the camera position.

With these multi-object 3D scenes composed at each timestep, we jointly render and optimize Gaussians from all objects using the RGB rendering loss $\mathcal{L}_{rgb}$ [18] and flow rendering loss $\mathcal{L}_{flow}$ [5, 47].

3.2.2. Instance and Occlusion-aware Gaussian Splatting

Jointly optimizing deformations can introduce a failure mode where Gaussians from different objects "jump" to another during close interactions or occlusions. To mitigate this, we assign a one-hot encoding instance label to each Gaussian as a fixed Gaussian attribute and incorporate an instance mask rendering loss.

We extend the differentiable 3D Gaussian renderer from [18] to render these instance labels, giving us an instance segmentation map for each frame. We then compare the rendered instance segmentation map $\hat{y}_k$ against the masks obtained from the pre-trained object mask trackers y_k (Section 3.2) using a standard negative log-likelihood loss $\mathcal{L}_{\text{class}} = -\sum_k y_k \log \hat{y}_k$.

During optimization, the one-hot class labels remain fixed, so any adjustments to the rendered instance segmentation map $\hat{y}_k$ can only result from Gaussian movements in the 3D space. This approach ensures that the deformed Gaussians do not undergo identity switches (i.e. "jump" to another object), allowing us to model complex multi-object occlusions accurately.

3.2.3. Robust Scene Modeling for In-the-Wild Videos

Long videos captured in natural environments often exhibit shadow and illumination changes as objects move, leading to variations in observed RGB values. To model this behavior, we draw on techniques from 3D scene reconstruction with unconstrained multiview image sets [19, 28] and diverge from previous dynamic Gaussian splatting works [27, 47], which keep Gaussian colors fixed over time. Instead, we allow Gaussian colors to change smoothly over time by introducing an additional MLP head on top of the deformation K-plane [4] to predict time-dependent adjustments to spherical harmonic (SH) coefficients in response to these observed color changes. The K-plane provides an inductive bias where Gaussians nearby in space and time should have similar color changes. An additional L1 regularization term $\mathcal{L}_{\text{reg}}$ is introduced on the magnitude of SH changes to prevent Gaussians from abusing this relaxation to explain all visual changes in the video.

3.2.4. Object-centric SDS Optimization in Global Scene

In monocular videos, unobserved views of the scene lack constraints, so we rely on view-conditioned generative models for regularization [33]. However, these view-conditioned generative models work best in object-centric settings, while GenMOJO directly represents all objects within a global coordinate frame. To bridge this gap, we employ affine transformations to transform object Gaussians from the world coordinate frame to object-centric coordinate frames and back, combining object-centric SDS and rendering losses in a unified framework.

Specifically, we fit bounding boxes to each object segmentation mask to obtain BBox_t in frame t. We then solve for the 2D affine warps $\{W_t\}$ that transforms BBox_t to



Figure 3. **Sample annotation from MOSE-PTS.** We visualize annotated points that are not occluded, with corresponding point tracks displayed in matching colors.

bounding boxes that correspond to that of an object-centric video. Geometrically, each warp W_t can be interpreted as a translation to reposition the object at the center of a sphere on which virtual cameras are sampled, and a scaling factor to resize it to the desired dimensions. W_t is then unprojected to 3D using the camera projection matrix and applied to object Gaussians to transform and render them in an object-centric frame. In this way, we can jointly optimize the object Gaussians using the SDS loss by rendering in the object-centric frame, and rendering losses on the observed video frames I_t under the reference camera pose.

3.2.5. Optimization Objective

Our optimization objective is described by a combination of $\mathcal{L}_{rgb}$ and $\mathcal{L}_{flow}$ (Section 3.2.1), the SDS loss, $\mathcal{L}_{class}$ (Section 3.2.2), and $\mathcal{L}_{reg}$ (Section 3.2.3). We also include the local rigidity loss [27] as an additional regularization term. Please refer to the supplementary materials for the weighting coefficients for each loss term.

4. MOSE-PTS Dataset and Analysis

The proposed MOSE-PTS dataset includes 20 videos selected from MOSE [6], a dataset designed for complex video object segmentation with substantial real-world occlusions. We filter the videos using 2D visible object masks to exclude low-interaction scenes, ensuring meaningful object interactions and tracking challenges. Compared to contemporary datasets like TAP-Vid-DAVIS[7], which often feature salient and isolated objects, MOSE-PTS presents more complex, cluttered scenes. Annotators labeled up to five objects per video, with five key points per object, and all annotations were conducted at a high resolution of 1080p to ensure precision and detail. Some annotation samples are visualized in Figure 3. Further annotation details can be found in the supplementary material.

Point Trajectory Analysis. In Figure 4, we compare MOSE-PTS with TAP-Vid-DAVIS [7] based on four metrics: average number of points, occlusion rate per trajectory, motion range, and occlusion rate per video. MOSE-PTS has a higher average number of points, requiring more simultaneous tracking and increasing the complexity of the dataset. Additionally, MOSE-PTS has more complex multi-object occlusions and a wider motion range than TAP-Vid-DAVIS.

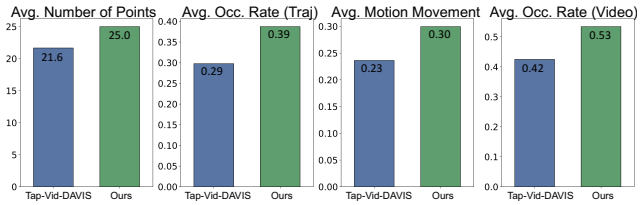


Figure 4. **Statistical comparison between MOSE-PTS and Tap-Vid-DAVIS [7].** Avg. number of points: the labeled points in each dataset. Avg. OCC. Rate (Traj / Video): the mean occlusion rate across individual trajectories, capturing the frequency of occlusions within distinct trajectories or videos. Avg. Motion Movement: the mean displacement of points between frames. These metrics show that MOSE-PTS contains more complex point tracks that have varied motion and undergo large occlusions.

5. Experiments

We evaluate GenMOJO for its ability to generate novel views of scenes and to track points in videos through occlusions. Our experiments aim to address these questions: **(1)** How does GenMOJO compare to other methods [5, 14, 37] in terms of novel view synthesis? **(2)** How does GenMOJO perform in tracking points through occlusions compared to existing state-of-the-art supervised point trackers [15, 16] and test-time optimization [5, 37, 47] methods? **(3)** To what extent do generative priors and object-based video decomposition contribute to improved view synthesis and point tracking? **(4)** Do object-centric generative priors [24] in GenMOJO outperform priors obtained from scene-level view-conditioned generative models [39, 40]?

Datasets. To evaluate GenMOJO on complex, real-world scenarios and provide a strong basis for comparison with existing baselines, we select a challenging subset of 15 multi-object videos from DAVIS [7, 32] and incorporate the 30 videos from our newly proposed MOSE-PTS dataset. This combined evaluation set spans a wide range of difficult scenes with diverse object motions and interactions, enabling a rigorous assessment of GenMOJO's ability to generate coherent 4D scenes that handle object interactions and accurately track points across occlusions.

Implementation Details. We conduct experiments on a 48GB A6000 GPU. We crop and scale the individual objects to around 65% of the image size for object lifting. For static 3D Gaussian optimization, we optimize for 1000 iterations with a batch size of 16, similar to previous works [5, 37]. To optimize the deformations, we optimize for 40 times the number of frames with a batch size of 8 and accumulate gradients for 2 steps. Details are provided in the appendix.



Figure 5. **Video to 4D Scene Generation Comparisons.** We render Shape of Motion [47], DreamScene4D [5], and GenMOJO on a DAVIS video (left) and a MOSE video (right) from the reference view (top three rows) and a novel view (bottom two rows). We can see that the baselines produce artifacts or incorrectly model spatial relationships between objects, while our method does not exhibit such errors. We also render novel views for DreamScene4D and our method, where we can similarly observe improvements.

5.1. Video to 4D Scene Generation

Baselines. We consider the following baselines for evaluating 4D scene generation: **(1)** Consistent4D [14], a method for video-to-4D generation from monocular videos that fits dynamic NeRFs per video using rendering losses and score distillation. **(2)** DreamGaussian4D [37], a 4D video generation method that, like our approach, uses dynamic Gaussian Splatting but does not employ object priors and thus lacks video decomposition. **(3)** DreamScene4D [5], a recent SOTA approach for 4D generation from multi-object monocular videos using dynamic Gaussian Splatting, making it the most directly comparable to our method. However, unlike GenMOJO, it optimizes each object's motion independently, whereas GenMOJO jointly optimizes motion across all objects to capture object interactions.

Evaluation Metrics. The quality of 4D generation can be assessed in two key aspects: the view rendering quality of the generated 3D geometry and the accuracy of the 3D motion. This section focuses on view rendering quality. Following previous works [14, 37], we report CLIP [35] and LPIPS [57] scores between four novel-view rendered frames and the input video frames, and calculate the average score per video. These metrics evaluate the semantic similarity between rendered and input video frames. Furthermore, we conducted a user study on MOSE videos, using a two-way voting method to compare GenMOJO with the state-of-the-art method DreamScene4D. Finally, we measure PSNR by comparing the rendered frame from the reference camera view to the input video sequence, providing a metric for the faithfulness of the 4D generation results.

4D Generation Results. The 4D generation results on MOSE and DAVIS videos are presented in Table 1. Since GenMOJO directly models objects in world space, it achieves the most faithful scene representation, as indicated

Table 1. **Video to 4D Scene Generation Comparisons**. We report the PSNR score for the re-rendered video under the reference camera pose, and the CLIP and LPIPS scores in novel views for MOSE [10] and DAVIS [32]. For user preference, we report the preference between DreamScene4D and GenMOJO in a two-way voting. GenMOJO outperforms existing methods for both novel view synthesis and produces faithful re-renderings in the reference camera view.

Method	MOSE				DAVIS		
	CLIP ↑	PSNR ↑	LPIPS ↓	User Pref.	CLIP ↑	PSNR ↑	LPIPS ↓
Consistent4D [14]	77.78	22.59	0.172	-	77.17	22.31	**0.146**
DreamGaussian4D [37]	81.96	17.82	0.195	-	81.62	17.68	0.183
DreamScene4D [5]	85.16	22.98	0.169	36.8%	84.13	21.73	0.163
GenMOJO (**Ours**)	**85.41**	**25.56**	**0.168**	**63.2%**	**84.17**	**23.51**	0.163

by its high PSNR scores on both MOSE and DAVIS. GenMOJO also produces high-quality novel views, supported by its CLIP and LPIPS scores, which measure semantic similarity with the reference view, and by the user preference study. Consistent4D and DreamGaussian4D, originally designed for object-centric inputs, struggle with multi-object scenes like those in MOSE and DAVIS. This limitation often leads to distorted 3D geometry or artifacts due to ineffective motion optimization. While DreamScene4D can handle multi-object videos, it optimizes each object independently, which limits its ability to model inter-object occlusions or interactions in more complex videos.

We show qualitative comparisons on MOSE and DAVIS videos between in Figure 5, highlighting instances where DreamScene4D fails to capture accurate relationships between interacting objects, leading to clipping problems. Additionally, we display reference view renderings from Shape of Motion, a recent Gaussian splatting-based method for point tracking. Since Shape of Motion only models visible parts of the scene, we exclude its results for novel views, and observe noisy renderings even in the reference view. These results demonstrate the robustness of GenMOJO in handling complex real-world videos with frequent occlusions and object interactions.

5.2. 4D Gaussian Motion Accuracy

Baselines. For evaluating Gaussian motion accuracy, we compare GenMOJO with DreamGaussian4D [37] and DreamScene4D [5], as in the previous section. Consistent4D [14] is not included since extracting accurate point motion from implicit NeRF representations is highly non-trivial and beyond our scope. We also include Shape of Motion [47], a recent optimization-based point-tracking method that employs 4D Gaussian Splatting. However, unlike GenMOJO, Shape of Motion does not leverage object or generative priors in optimizing dynamic Gaussians.

Evaluation Metrics. To evaluate motion accuracy, we use both Average Trajectory Error (ATE) and Median Trajectory Error (MTE) [58], which measure the L2 distance between the estimated tracks and the ground truth tracks in all timesteps. MTE is included because it is less sensitive to extreme failure cases, providing a robust assessment of typical tracking accuracy. We also report the Average End Point Error (A-EPE) and Median End Point Error (M-EPE) [5, 11], which calculate the L2 distance specifically at the final timestep. This provides specific insights into the tracking accuracy over long time horizons. For consistency, the predicted point tracks of all the methods are normalized to a resolution of 256×256 during the evaluation [7, 15].

Point Tracking Results. Table 2 presents the point tracking results on MOSE and DAVIS video datasets, including comparisons with SOTA supervised methods [15, 16] trained on point annotations serving as an upper bound for performance. Supervised trackers like CoTrackerV2 and CoTrackerV3 show significantly lower accuracy on MOSE compared to DAVIS, highlighting the greater challenge posed by MOSE videos due to frequent, prolonged occlusions and larger object and point displacements. GenMOJO achieves the lowest tracking error in all semi-supervised and optimization-based methods, excelling in each evaluation metric. It also performs competitively on the simpler DAVIS dataset.

Both DreamScene4D [5] and GenMOJO benefit from object priors that help the Gaussians stay anchored to the same object throughout the video, enhancing tracking accuracy over other baselines, where tracked points often drift and get lost on different objects. This drift is a common issue for GMRW [42], Shape-of-Motion [47], and CoTracker [15], especially during occlusions. However, DreamScene4D, unlike GenMOJO, lacks cross-object interaction modeling, which hinders its performance during occlusions. Without joint motion optimization, DreamScene4D's Gaussian motion tracks in separate objects struggle to account for observed interactions in the video, causing tracks to drift or flicker within objects during occlusions, ultimately reducing tracking accuracy.

5.3. Ablation Experiments

We perform ablation studies to assess the impact of various components in GenMOJO on the generation quality and the motion accuracy, as shown in Table 3 using MOSE videos. First, we evaluate the effect of removing joint splatting

Table 2. **Point tracking comparison between 4D generation / reconstruction methods.** We report the Average Trajectory Error (ATE), Median Trajectory Error (MTE) [58], Average End Point Error (A-EPE), and Median End Point Error (M-EPE) [5] in MOSE [10] and DAVIS [7, 32]. GenMOJO outperforms existing methods that were not trained on point tracks in MOSE, and achieves competitive performance when compared against learning-based methods.

Method	MOSE				DAVIS			
	ATE ↓	MTE ↓	A-EPE↓	M-EPE↓	ATE ↓	MTE ↓	A-EPE↓	M-EPE↓
(a) ***Not trained on point tracking data***								
GMRW [42]	24.85	20.87	42.59	36.99	42.17	26.73	75.61	60.14
DreamGaussian4D [37]	27.73	25.71	46.18	43.55	30.27	28.18	61.5	55.69
Shape of Motion [47]	18.39	10.40	28.25	18.30	8.91	**3.93**	18.94	**6.08**
DreamScene4D [5]	13.47	10.49	22.48	16.97	9.59	6.50	17.47	10.29
GenMOJO (**Ours**)	**9.91**	**8.47**	**15.46**	**11.11**	**8.16**	6.32	**14.14**	9.36
(b) ***Trained on point tracking data***								
CoTrackerv2 [15]	15.74	12.15	27.96	20.35	21.90	2.08	26.22	3.49
CoTrackerv3 [16]	11.54	8.33	22.16	16.11	21.24	1.74	24.0	3.11

Table 3. **Ablation experiments**. We report motion accuracy and view synthesis metrics in MOSE [10] by removing different proposed components in GenMOJO. We additionally replace object-centric SDS with scene-level SDS [40], and replace DepthCrafter with DepthAnything-V2 [52] for depth initialization.

Method	MOSE			
	A-EPE ↓	M-EPE ↓	CLIP ↑	PSNR ↑
Full Model	**15.46**	**11.11**	**85.41**	**25.56**
w/o Multi-object Joint Splatting	23.48	16.97	85.18	22.88
w/o Transformed Obj-centric rendering	19.44	14.50	81.44	25.30
w/o Instance Masks Rendering	16.77	12.49	84.88	25.19
Scene-Level SDS [40]	20.03	15.49	83.05	25.17
DepthAnything-V2 [52]	17.32	12.84	85.29	25.33

optimization and instead use a method similar to DreamScene4D, where objects are optimized independently. This change leads to a marked drop in motion accuracy and PSNR, confirming that joint splatting is crucial for accurately modeling object occlusions and interactions.

Next, we investigate the role of unseen object modeling by testing the integration of SDS during motion optimization. As expected, omitting this results in a substantial decrease in the quality of novel view synthesis. The motion accuracy also suffers, which we hypothesize is due to Gaussians in unobserved views becoming underconstrained without SDS and can move unpredictably. We further ablate with a scene-level SDS model [40] rather than object-centric SDS ones [24], which also led to performance drops. This may be due to object-centric NVS diffusion models outperforming scene-level ones, and leading to lower quality SDS gradients.

To examine GenMOJO's robustness to different depth models, we replace DepthCrafter [13] with DepthAnything-V2 [52], which generates less consistent depth across frames. We observe a minor performance impact, likely due to occasional depth inaccuracies. Lastly, we remove the instance mask rendering loss and find a reduction in motion accuracy, though the novel view synthesis quality remains largely unchanged. We observed that this is due to some Gaussians drifting between objects without this loss, leading to less accurate motion tracking.

5.4. Limitations

Despite the progress demonstrated in this paper, GenMOJO has several limitations: (1) Similar to previous methods [5, 37], our approach relies on the generalization capabilities of the view-conditioned generative model, which struggles with videos featuring unconventional camera poses. (2) Although our joint motion optimization procedure can correctly determine the depth ordering of objects, jittery depth estimations can still cause objects to "jump" along the depth axis. (3) Due to the reliance on test-time optimization, our method is not suitable for online applications. We hope to develop novel feedforward models that can enable both real-time inference and provide stronger priors for test-time optimization in future work.

6. Conclusion

In this work, we introduce GenMOJO, a novel approach for generating complete 4D scenes from complex monocular videos containing multiple objects, prolonged occlusions, and intricate inter-object interactions. GenMOJO leverages state-of-the-art pre-trained models to decompose scenes into individual object trajectories while employing a joint motion optimization strategy to capture inter-object dynamics accurately. Experimental results in in-the-wild video datasets, such as DAVIS and MOSE, highlight the superiority of our method in view-synthesis tasks, outperforming existing baselines. Furthermore, we annotated the point tracks in MOSE videos to assess the accuracy of Gaussian motion, revealing that the Gaussian tracks generated by GenMOJO outperform previous methods in challenging scenarios characterized by large movements and extended occlusions.

References

[1] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. *arXiv preprint arXiv:2311.17984*, 2023. 2

[2] Sherwin Bahmani, Xian Liu, Yifan Wang, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. *arXiv preprint arXiv:2403.17920*, 2024. 2

[3] Marcel Büsching, Josef Bengtson, David Nilsson, and Mårten Björkman. Flowibr: Leveraging pre-training for efficient neural image-based rendering of dynamic scenes. *arXiv preprint arXiv:2309.05418*, 2023. 2

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, 2023. 2, 4

[5] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. In *NeurIPS*, 2024. 1, 2, 3, 4, 5, 6, 7, 8

[6] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip HS Torr, and Song Bai. Mose: A new dataset for video object segmentation in complex scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20224–20234, 2023. 2, 3, 5

[7] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. In *NeurIPS*, 2022. 5, 7, 8

[8] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 2, 3

[9] Quankai Gao, Qiangeng Xu, Zhe Cao, Ben Mildenhall, Wenchao Ma, Le Chen, Danhang Tang, and Ulrich Neumann. Gaussianflow: Splatting gaussian dynamics for 4d content creation. *arXiv preprint arXiv:2403.12365*, 2024. 3

[10] Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, et al. Kubric: A scalable dataset generator. In *CVPR*, 2022. 7, 8

[11] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *European Conference on Computer Vision*, pages 59–75. Springer, 2022. 7

[12] Eric Heiden, Ziang Liu, Vibhav Vineet, Erwin Coumans, and Gaurav S Sukhatme. Inferring articulated rigid body dynamics from rgbd video. In *IROS*, 2022. 1

[13] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 3, 4, 8

[14] Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. In *ICLR*, 2024. 1, 2, 3, 5, 6, 7

[15] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 3, 5, 7, 8

[16] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 5, 7, 8

[17] Lei Ke, Mingqiao Ye, Martin Danelljan, Yifan Liu, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Segment anything in high quality. In *NeurIPS*, 2023. 3

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4):1–14, 2023. 2, 3, 4

[19] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *NeurIPS*, 2024. 4

[20] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2

[21] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022. 2

[22] Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. *arXiv preprint arXiv:2312.13763*, 2023. 2

[23] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[24] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 1, 3, 5, 8

[25] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *CVPR*, 2023. 2

[26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 2

[27] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 2, 3, 4, 5

[28] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 4

[29] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

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

[31] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 2

[32] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 5, 7, 8

[33] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 3, 4

[34] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 6

[36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 4

[37] Jiawei Ren, Liang Pan, Jiaxiang Tang, Chi Zhang, Ang Cao, Gang Zeng, and Ziwei Liu. Dreamgaussian4d: Generative 4d gaussian splatting. *arXiv preprint arXiv:2312.17142*, 2023. 2, 3, 5, 6, 7, 8

[38] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 3, 4

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 5

[40] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, et al. Zeronvs: Zero-shot 360-degree view synthesis from a single real image. In *CVPR*, 2024. 1, 5, 8

[41] Jenny Seidenschwarz, Qunjie Zhou, Bardienus Duisterhof, Deva Ramanan, and Laura Leal-Taixé. Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. *arXiv preprint arXiv:2409.02104*, 2024. 3

[42] Ayush Shrivastava and Andrew Owens. Self-supervised any-point tracking by contrastive random walks. *ECCV*, 2024. 7, 8

[43] Uriel Singer, Shelly Sheynin, Adam Polyak, Oron Ashual, Iurii Makarov, Filippos Kokkinos, Naman Goyal, Andrea Vedaldi, Devi Parikh, Justin Johnson, et al. Text-to-4d dynamic scene generation. *arXiv preprint arXiv:2301.11280*, 2023. 2

[44] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. *arXiv preprint arXiv:2406.18717*, 2024. 2

[45] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *ICLR*, 2024. 3

[46] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19795–19806, 2023. 3

[47] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 2, 3, 4, 5, 6, 7, 8

[48] Shizun Wang, Xingyi Yang, Qiuhong Shen, Zhenxiang Jiang, and Xinchao Wang. Gflow: Recovering 4d world from monocular video. *arXiv preprint arXiv:2405.18426*, 2024. 2

[49] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 2, 3

[50] Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *CVPR*, 2024. 1

[51] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 3

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. 8

[53] Qitong Yang, Mingtao Feng, Zijie Wu, Shijie Sun, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Beyond skeletons: Integrative latent mapping for coherent 4d sequence generation. *arXiv preprint arXiv:2403.13238*, 2024. 2

[54] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2

[55] Yuyang Yin, Dejia Xu, Zhangyang Wang, Yao Zhao, and Yunchao Wei. 4dgen: Grounded 4d content generation with spatial-temporal consistency. *arXiv preprint arXiv:2312.17225*, 2023. 1, 3

[56] Yifei Zeng, Yanqin Jiang, Siyu Zhu, Yuanxun Lu, Youtian Lin, Hao Zhu, Weiming Hu, Xun Cao, and Yao Yao. Stag4d: Spatial-temporal anchored generative 4d gaussians. *arXiv preprint arXiv:2403.14939*, 2024. 1, 3

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6

[58] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *ICCV*, 2023. 7, 8

[59] Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. *arXiv preprint arXiv:2311.16854*, 2023. 2