Orientation-anchored Hyper-Gaussian for 4D Reconstruction from Casual Videos

Junyi Wu[1] **Jiachen Tao**[1] **Haoxuan Wang**[1]
Gaowen Liu[2] **Ramana Rao Kompella**[2] **Yan Yan**[1]*
[1]University of Illinois Chicago [2]Cisco Research
https://github.com/adreamwu/OriGS

Abstract

We present **Orientation-anchored Gaussian Splatting** (**OriGS**), a novel framework for high-quality 4D reconstruction from casually captured monocular videos. While recent advances extend 3D Gaussian Splatting to dynamic scenes via various motion anchors, such as graph nodes or spline control points, they often rely on low-rank assumptions and fall short in modeling complex, region-specific deformations inherent to unconstrained dynamics. OriGS addresses this by introducing a hyperdimensional representation grounded in scene orientation. We first estimate a **Global Orientation Field** that propagates principal forward directions across space and time, serving as stable structural guidance for dynamic modeling. Built upon this, we propose **Orientation-aware Hyper-Gaussian**, a unified formulation that embeds time, space, geometry, and orientation into a coherent probabilistic state. This enables inferring region-specific deformation through principled conditioned slicing, adaptively capturing diverse local dynamics in alignment with global motion intent. Experiments demonstrate the superior reconstruction fidelity of OriGS over mainstream methods in challenging real-world dynamic scenes.

1 Introduction

Videos offer a window into the continuous flow of forms, light, and transformations that weave our experience of space and time. Among various sources, *casually captured monocular videos* stand as the most ubiquitous and accessible. These videos are typically recorded by smartphones, handheld cameras, or consumer drones, spanning a broad spectrum of everyday scenes with diverse object compositions and unconstrained motion patterns. Recovering the underlying dynamic reality from such inputs remains a foundational pursuit in computer graphics and 3D vision [33], boosting applications in autonomous driving [72, 21], robotics [30, 53], and virtual/augmented reality [81, 10].

At its core, 4D reconstruction aims to recover the time-varying geometry, appearance, and motion of real-world environments from videos [82], bridging visual observations with underlying scene dynamics. 3D Gaussian Splatting (3DGS) [37] has recently emerged as a powerful solution for static scene reconstruction, offering a compact and expressive point-based representation that enables high-fidelity modeling and real-time rendering. Building on its success, substantial efforts have been devoted to extending 3DGS to dynamic scenes [49, 76, 71, 64]. These methods primarily capture temporal evolution by applying per-Gaussian deformations, such as translation and rotation, across time. Recent works are further pursuing robust reconstruction from casually captured videos by modeling deformation through various motion anchors. Representative approaches include learning motion bases and weighting coefficients [67], constructing discrete 3D node graphs [48, 39, 44], and employing continuous cubic Hermite splines with sparse control points [77, 54].

Yet, reconstructing scene dynamics from casually captured monocular videos remains challenging due to the complexity of local motion. Regions within the scene often exhibit diverse motion

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

patterns, influenced by object articulation, interaction context, or progression through different action stages [61]. Although recent approaches incorporate explicit, structured deformation fields, they largely reduce complex dynamics to low-rank motion anchors, assuming that neighboring regions exhibit similar motion. While effective for smooth or rigid transformations, such deformation anchors are inherently fragile under unconstrained dynamics in real-world scenarios, where motion patterns may vary significantly across regions. Consequently, these formulations often struggle to capture complex local dynamics, resulting in spatial drift, structural fragmentation, or temporal inconsistency.

To address these challenges, we are motivated by the perspective that the diversity of local motion often reflects more than just positional or geometric variation; it suggests an underlying regularity in how different regions evolve over time. Intuitively, local deformations are not solely determined by spatial position or object shape, but also by how each part of a scene is *expected* to move within a broader dynamic context. Fortunately, while casual monocular videos lack explicit multi-view supervision, they do preserve object-centric motion cues that implicitly reflect these region-specific motion tendencies. Among them, we focus our exploration on a natural and grounded signal: *orientation*. In the physical world, while absolute position is sensitive to noise and scale ambiguity, orientation evolves under the smoother governance of local angular momentum and inertia [16, 27]. This makes it a more stable proxy for how motion is expected to unfold in complex dynamic scenes, indicating the forward tendencies across regions. In this work, we embrace orientation as a dynamic anchor for 4D reconstruction. Our key insight is to treat orientation as a structural signal that organizes how different parts of the scene deform and interact, thereby capturing coherent evolution.

Building on this perspective, we introduce **Orientation-anchored Gaussian Splatting** (**OriGS**), a novel framework for high-quality 4D reconstruction from casual monocular videos. Our OriGS comprises two synergistic components for dynamic modeling. ❶ We first construct a **Global Orientation Field** that captures long-range scene evolution by estimating and propagating principal forward directions over time via localized structure alignment. This orientation field provides stable oriented anchors that extend across space and time, which help organize diverse motion patterns in different regions. ❷ On top of this structured foundation, we propose **Orientation-aware Hyper-Gaussian**, a hyperdimensional representation to model complex local dynamics throughout the scene. We associate each Gaussian primitive with a unified probabilistic state defined over a structured manifold of time, space, geometry, and orientation. By conditioning this representation on the scene's evolving orientation field, we infer region-specific deformations through a principled conditioned slicing strategy. This allows OriGS to adaptively model how different parts of the scene deform over time and how their motions evolve in alignment with the underlying dynamic intent. Together, these two components enable OriGS to robustly reconstruct dynamic scenes from casual videos, bridging global coherence and local motion diversity in a unified and principled manner.

We validate the effectiveness of OriGS on a range of casual monocular videos, including DAVIS [59], OpenAI SORA [6], YouTube-VOS [74], and the DyCheck benchmark [19]. OriGS consistently recovers sharper geometry and more coherent motion compared to recent state-of-the-art methods, demonstrating its superior reconstruction fidelity in real-world dynamic scenes.

2 Related Works

Dynamic 3D Representations. Dynamic scene modeling has been widely studied by extending static 3D representations with deformation mechanisms [20, 9]. Neural Radiance Fields (NeRF) [51] represent scenes as volumetric functions, and are adapted to dynamic settings by learning deformation fields that warp canonical points over time [55, 60, 56, 17, 34, 70]. Later works introduce temporal embeddings [57, 40], or hybrid representations for faster rendering [2, 18, 8, 62]. However, their implicit volumetric nature limits real-time inference and structure-aware modeling. Recently, 3D Gaussian Splatting (3DGS) [37] has emerged as a powerful alternative for scene representation [4, 82]. Dynamic extensions incorporate deformation fields [49, 71, 76, 15, 3, 31], apply trajectory interpolation [41, 38, 77], or directly embed time into Gaussian primitives [14, 75]. Yet, lacking explicit guidance from underlying structures, these methods are generally incapable of faithfully recovering complex motion in real-world dynamic scenes.

Hyperdimensional Gaussians. Recent work has enhanced the expressiveness of Gaussian representations by embedding auxiliary signals into higher-dimensional spaces. 4DGS [75, 14] incorporates time to model spatio-temporal volumes, while N-DG [11] generalizes to a latent space embedding po-

sition, view direction, and material cues. To capture view-dependent effects, 6DGS [22, 24] integrates angular information and modulates opacity and color accordingly, and 7DGS [23] further extends this with temporal modeling. In this work, we explore hyperdimensional Gaussians as structured probabilistic states, modeling space, time, geometry, and orientation in a unified representation.

4D Reconstruction from Casual Monocular Videos. Recovering dynamic representations from casually captured monocular videos is highly ill-posed. Early efforts enhance NeRFs by specialized deformation fields [60, 55, 56], long-range feature aggregation [43], or point trajectory prediction [66]. While accounting for scene changes, they predominantly consider controlled settings with teleporting cameras or quasi-static scenes [19]. Recent advances shift toward Gaussian-based representations, leveraging 2D priors to enable reconstruction in the wild. Methods like Shape-of-Motion [67], Gaussian Marbles [64], MoSca [39], and MoDGS [46] guide per-Gaussian deformation through various motion anchors. Others adopt compact motion models, such as spline interpolation [54] or hierarchical structures [44], to trace temporal trajectories. Nonetheless, their reliance on low-rank motion priors limits the capacity to capture the complex, spatially varying dynamics present in real-world environments. In contrast, we propose to organize deformation within a global orientation field and condition dynamic modeling on a unified hyperdimensional representation, enabling coherent reconstruction under diverse motion patterns.

3 Preliminaries

3D Gaussian Splatting. 3DGS [37] represents scenes as a set of explicit 3D Gaussian primitives, each specified by a spatial center $\boldsymbol{\mu}_\mathbf{p} \in \mathbb{R}^3$ and an anisotropic covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{p})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\mathbf{p})\right), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^3$ denotes a 3D spatial location. To ensure positive semi-definiteness, $\boldsymbol{\Sigma}$ is factorized as $\boldsymbol{\Sigma} = \mathbf{RSS}^\top\mathbf{R}^\top$ in practice, where $\mathbf{S}$ is a diagonal scaling matrix and $\mathbf{R}$ is a rotation matrix aligning the Gaussian with the global coordinate system. Moreover, each 3D Gaussian associates an opacity value $\sigma \in \mathbb{R}$ and a color vector $\mathbf{c} \in \mathbb{R}^3$. During rendering, contributions from all Gaussians overlapping a pixel are composited via alpha blending to provide the final color.

In this work, we adopt 3DGS as a differentiable and compact representation for dynamic scenes, which can be efficiently rendered into images with a rasterization pipeline.

3D Trajectories from Monocular Videos. Given a monocular video of T frames, our first step toward 4D reconstruction is recovering long-range 3D trajectories of scene points over time. Each trajectory is defined as a sequence $\{\boldsymbol{\tau}^t\}_{t=1}^T$, where $\boldsymbol{\tau}^t \in \mathbb{R}^3$ denotes the 3D position at frame t. However, estimating such trajectories from monocular inputs is inherently ill-posed due to depth ambiguity from 3D-to-2D projection. To mitigate this, we combine metric depth estimation [28, 29, 58] with long-range pixel tracking [12, 36, 73]. We predict per-frame depth maps $d^t : \mathbb{R}^2 \to \mathbb{R}_+$ and extract 2D trajectory $\{\mathbf{u}^t\}_{t=1}^T$, where $\mathbf{u}^t \in \mathbb{R}^2$ denotes the pixel location at frame t. Consistent with common practices [64, 73, 39], we then lift each 2D trajectory into 3D world space, aided by the depth priors:

$$\boldsymbol{\tau}^t = \mathbf{W}^t \pi_\mathbf{K}^{-1}\left(\mathbf{u}^t, d^t(\mathbf{u}^t)\right), \tag{2}$$

where $\pi_\mathbf{K}(\cdot)$ denotes the projection function from camera to image space defined by intrinsics $\mathbf{K}$, and $\mathbf{W}^t$ is the camera pose at frame t. When camera parameters are unavailable, we estimate them via bundle adjustment [45, 39].

The resulting sequence $\{\boldsymbol{\tau}^t\}_{t=1}^T$ provides a 3D trajectory across the scene. We further extract principal forward directions from these trajectories to construct oriented anchors, which provide structured and temporally coherent guidance for dynamic modeling in our framework.

4 Method

We introduce **Orientation-anchored Gaussian Splatting** (**OriGS**), a novel framework for 4D reconstruction from casual videos. OriGS is built upon two synergistic components: **(i)** a **Global Orientation Field** (Section 4.1) that captures long-range scene evolution via propagated orientation cues; and **(ii)** an **Orientation-aware Hyper-Gaussian** (Section 4.2) that models complex region-specific dynamics via orientation-conditioned inference in a unified hyperdimensional representation.

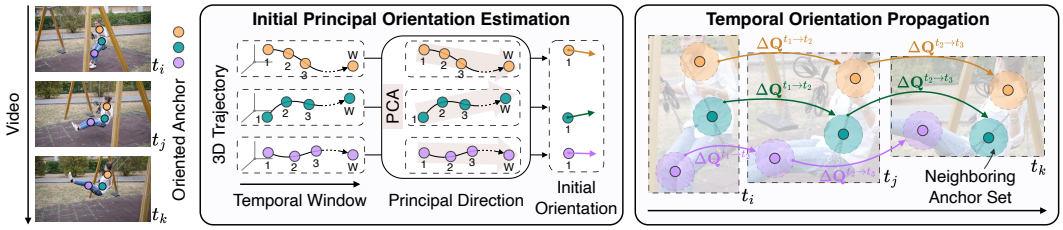

Figure 1: **Illustration of Global Orientation Field.** Given 3D trajectories recovered from monocular video, we extract a structured set of oriented anchors. **(Left)** Anchors' positions are initialized along tracked points across frames. **(Middle)** We estimate initial principal orientations by applying PCA over short temporal windows, reflecting the dominant forward direction per anchor. **(Right)** These initial orientations are then propagated over time via localized Procrustes alignment, producing a temporally coherent orientation field across the scene.

4.1 Global Orientation Field

Initial Principal Orientation Estimation. Given the recovered 3D trajectory $\{\boldsymbol{\tau}^t\}_{t=1}^T$ of scene points (Section 3), we convert these raw translation paths into a structured set of *oriented anchors*, each carrying a spatial position and an associated local orientation. These anchors serve as the basis for constructing our Global Orientation Field.

To initialize orientation, we estimate the principal direction of each anchor by analyzing the dominant motion in the early portion of its trajectory. Concretely, we unfold each trajectory and extract the first W frames as a temporal window, forming $\{\boldsymbol{\tau}^t\}_{t=1}^W$. We center the points using the temporal mean $\bar{\boldsymbol{\tau}} = \frac{1}{W}\sum_{t=1}^W \boldsymbol{\tau}^t$, and define $\hat{\boldsymbol{\tau}}^t = \boldsymbol{\tau}^t - \bar{\boldsymbol{\tau}}$. We then compute the covariance matrix of the centered points and apply principal component analysis (PCA) to extract the dominant direction:

$$\text{PCA}(\mathbf{C}) = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3], \quad \text{where} \quad \mathbf{C} = \frac{1}{W}\sum_{t=1}^{W} \hat{\boldsymbol{\tau}}^t \hat{\boldsymbol{\tau}}^{t\top}. \tag{3}$$

The leading eigenvector $\mathbf{v}_1$ reflects the dominant motion axis, which we assign as the anchor's forward direction to construct its initial local orientation $\mathbf{O}^1 \in SO(3)$.

Temporal Orientation Propagation. To extend the initial principal orientation across time, we estimate a time-varying orientation sequence $\{\mathbf{O}^t\}_{t=1}^T$ for each anchor, capturing its evolving motion direction throughout the scene. This can be formulated as a Procrustes alignment problem [26, 25]. At each time step t, we identify a local neighborhood $\{\boldsymbol{\tau}_k^t\}_{k=1}^K$ around each anchor based on Euclidean distance and align it to its reference configuration by optimizing:

$$\min_{\mathbf{T}^t \in SO(3)} \sum_{k=1}^{K} \left\| (\boldsymbol{\tau}_k^t - \bar{\boldsymbol{\tau}}^t) - \mathbf{T}^t(\boldsymbol{\tau}_k^1 - \bar{\boldsymbol{\tau}}^1) \right\|^2, \tag{4}$$

where $\bar{\boldsymbol{\tau}}^t$ and $\bar{\boldsymbol{\tau}}^1$ are spatial centroids of the anchor set, and $\mathbf{T}^t$ represents the rotation transformation.

This alignment is solved by applying singular value decomposition (SVD) to the covariance matrix of the localized anchor set. The optimal rotation transformation is then given by $\mathbf{T}^{t*} = \mathbf{V}\mathbf{U}^\top$, where $\mathbf{V}$ and $\mathbf{U}$ are obtained from SVD. As a correction step, we adjust the last column based on the sign of its determinant. Finally, we compute the anchor's orientation $\mathbf{O}^t$ by composing the optimal transformation with the initial principal orientation:

$$\mathbf{O}^t = \mathbf{T}^{t*} \cdot \mathbf{O}^1. \tag{5}$$

Repeating this process yields temporally coherent orientations $\{\mathbf{O}^t\}_{t=1}^T$. These propagated orientations constitute our Global Orientation Field, which encodes motion tendencies across the scene and provides stable anchors for modeling complex dynamics. We illustrate this process in Figure 1, where $\Delta\mathbf{Q}_i^{t\to t'} = (\mathbf{O}_i^{t'}, \boldsymbol{\tau}_i^{t'})(\mathbf{O}_i^t, \boldsymbol{\tau}_i^t)^{-1}$ denotes the relative anchor transformation in $SE(3)$ across time.

4.2 Orientation-aware Hyper-Gaussian

Dynamic scenes often exhibit complex, region-specific motion that cannot be fully explained by position or geometry alone. While casual monocular videos lack multi-view supervision, they retain

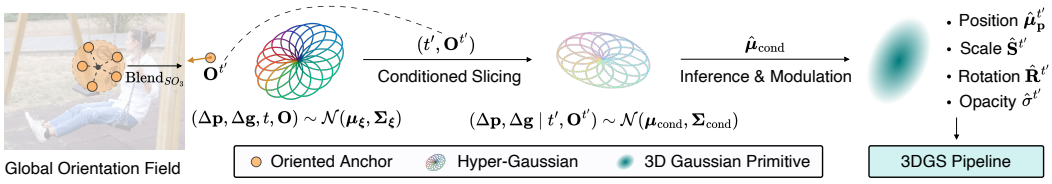

Figure 2: **Illustration of Orientation-aware Hyper-Gaussian.** We model complex dynamics using high-dimensional Gaussians defined over a structured manifold spanning time, space, geometry, and orientation. Given a target time t' and its local orientation $\mathbf{O}^{t'}$ (interpolated from the Global Orientation Field via $SO(3)$ blending), we perform conditioned slicing on the hyper-Gaussian to infer the expected dynamic states. The predicted $\hat{\boldsymbol{\mu}}_{\text{cond}}$ modulates 3D Gaussian primitives, which is then compatible with the 3DGS pipeline.

coherent orientation cues that implicitly reveal how different regions are expected to move. We leverage this structural signal and introduce **Orientation-aware Hyper-Gaussian**, a hyperdimensional representation that models local dynamics over a unified manifold spanning time, space, geometry, and orientation. By conditioning on the evolving orientation field, this formulation facilitates principled deformation inference aligned with the dynamic intent.

Multivariate Modeling of Dynamic States. To reflect complex motion patterns across space and time, we represent each region of the scene with a multivariate dynamic state that jointly encodes temporal progression and structural changes. Specifically, let $\mathbf{p}^t \in \mathbb{R}^3$ denote the position of a point at time t. We define its local dynamic state as:

$$\boldsymbol{\xi} = (\Delta\mathbf{p}, \Delta\mathbf{g}, t, \mathbf{O}) \in \mathcal{M}, \tag{6}$$

where $\Delta\mathbf{p}$ and $\Delta\mathbf{g}$ represent deviations in position and geometry (*e.g.*, scale and rotation), $\mathbf{O} \in SO(3)$ is the orientation derived from our Global Orientation Field, and $\mathcal{M}$ denotes a manifold spanning time, space, geometry, and orientation.

This formulation captures the intrinsic interdependence among these factors in dynamic scenes, treating them as jointly evolving rather than isolated variables. To model their correlations in a unified and flexible manner, we then represent each dynamic state $\boldsymbol{\xi}$ as a probabilistic entity with a multivariate Gaussian over $\mathcal{M}$:

$$\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}}), \quad \boldsymbol{\mu}_{\boldsymbol{\xi}} \in \mathcal{M}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\xi}} \in \mathbb{R}^{D\times D}, \tag{7}$$

where $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ is the canonical dynamic state, and $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ captures intra- and inter-correlations. By grounding this probabilistic state in the evolving global orientation field, our model is able to learn region-specific variability that aligns with the scene's underlying dynamic intent.

Hyper-Gaussian Representation. To realize our formulation, we build upon 3D Gaussian Splatting [37] and equip each primitive with the multivariate Gaussian distribution defined in Eq. (7). This results in a dynamic representation that encodes the localized motion and geometric variability within a unified probabilistic entity.

Specifically, we begin by initializing a set of 3D Gaussians $\{\mathcal{G}_j\}$ across the scene, each parameterized by its position $\boldsymbol{\mu}_{\mathbf{p}_j}$, rotation $\mathbf{R}_j$, scale $\mathbf{S}_j$, opacity σ_j, and color $\mathbf{c}_j$ (Section 3). To enable dynamic modeling, we associate each $\mathcal{G}_j$ with a state mean $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$:

$$\boldsymbol{\mu}_{\boldsymbol{\xi}} = (\boldsymbol{\mu}_{\Delta\mathbf{p}}, \boldsymbol{\mu}_{\Delta\mathbf{g}}, \mu_t, \boldsymbol{\mu}_{\mathbf{O}}), \quad \boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \begin{bmatrix} \boldsymbol{\Sigma}_{(\Delta\mathbf{p},\Delta\mathbf{g})} & \boldsymbol{\Sigma}_{(\Delta\mathbf{p},\Delta\mathbf{g}),(t,\mathbf{O})} \\ \boldsymbol{\Sigma}^{\top}_{(\Delta\mathbf{p},\Delta\mathbf{g}),(t,\mathbf{O})} & \boldsymbol{\Sigma}_{(t,\mathbf{O})} \end{bmatrix}, \tag{8}$$

where $\boldsymbol{\mu}_{\Delta\mathbf{p}} \in \mathbb{R}^3$ represents the spatial offset, $\boldsymbol{\mu}_{\Delta\mathbf{g}} \in \mathbb{R}^3_+ \times SO(3)$ describes local variations in scale and rotation, $\mu_t \in \mathbb{R}$ is the temporal coordinate, and $\boldsymbol{\mu}_{\mathbf{O}} \in SO(3)$ represents the orientation within the global field. The covariance matrix captures both intra- (*e.g.,* temporal or geometric) and cross-group (*e.g.,* spatial-orientational) dependencies to model complex dynamics.

For numerical stability, we parameterize $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$ via Cholesky decomposition [11, 22, 23]: $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathbf{L}_{\boldsymbol{\xi}}\mathbf{L}_{\boldsymbol{\xi}}^{\top}$, where $\mathbf{L}_{\boldsymbol{\xi}}$ is a lower-triangular matrix. Both $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and $\mathbf{L}_{\boldsymbol{\xi}}$ are treated as learnable parameters and jointly optimized with other Gaussian attributes. More details are provided in the supplementary materials.

Conditioned Slicing for Dynamic State Inference. Given the hyper-Gaussian representation, we seek to infer concrete dynamic behavior by conditioning on local motion context. Our core intuition

is that each part of the scene undergoes diverse localized deformation in response to how the region is expected to evolve under temporal and orientational cues.

Specifically, we define a query $(t', \mathbf{O}^{t'}) \in \mathbb{R} \times SO(3)$, representing the target time and its associated orientation. Given the probability $P(\boldsymbol{\xi}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\xi}}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}})$, we derive the conditional distribution over the spatial and geometric subspace:

$$P(\Delta\mathbf{p}, \Delta\mathbf{g} \mid t', \mathbf{O}^{t'}) = \mathcal{N}(\boldsymbol{\mu}_{\text{cond}}, \boldsymbol{\Sigma}_{\text{cond}}), \tag{9}$$

with conditional mean and covariance given by Gaussian slicing rules:

$$\boldsymbol{\mu}_{\text{cond}} = (\boldsymbol{\mu}_{\Delta\mathbf{p}}, \boldsymbol{\mu}_{\Delta\mathbf{g}}) + \boldsymbol{\Sigma}_{(\Delta\mathbf{p},\Delta\mathbf{g}),(t,\mathbf{O})}\boldsymbol{\Sigma}^{-1}_{(t,\mathbf{O})}(t' - \mu_t, \mathbf{O}^{t'} \ominus \boldsymbol{\mu}_{\mathbf{O}}), \tag{10}$$

$$\boldsymbol{\Sigma}_{\text{cond}} = \boldsymbol{\Sigma}_{(\Delta\mathbf{p},\Delta\mathbf{g})} - \boldsymbol{\Sigma}_{(\Delta\mathbf{p},\Delta\mathbf{g}),(t,\mathbf{O})}\boldsymbol{\Sigma}^{-1}_{(t,\mathbf{O})}\boldsymbol{\Sigma}_{(t,\mathbf{O}),(\Delta\mathbf{p},\Delta\mathbf{g})}, \tag{11}$$

where $\ominus$ denotes the relative rotation operator in $SO(3)$. Practically, rather than sampling $(\Delta\mathbf{p}, \Delta\mathbf{g})$ from the conditional distribution, which may introduce optimization instability, we adopt the maximum-likelihood estimation principle and use the conditional mean $\boldsymbol{\mu}_{\text{cond}}$ as the deterministic prediction:

$$\hat{\boldsymbol{\mu}}_{\text{cond}} := \mathbb{E}[\Delta\mathbf{p}, \Delta\mathbf{g} \mid t', \mathbf{O}^{t'}] = \boldsymbol{\mu}_{\text{cond}}. \tag{12}$$

This inferred local dynamic state captures the expected position and geometry variations given temporal and orientation contexts. Through this slicing process, the hyper-Gaussian can be adaptively projected onto various scene regions, enabling temporally and spatially coherent reconstruction.

Modulation of Hyper-Gaussian Attributes. To integrate our dynamic modeling into the efficient and differentiable rendering pipeline, we modulate 3D Gaussian primitives based on the inferred dynamic state. For a target time t' and local orientation $\mathbf{O}^{t'}$, we can obtain $\hat{\boldsymbol{\mu}}_{\text{cond}}$ from the conditioned distribution using Eq. (12) and apply it to the 3D Gaussian parameters accordingly:

$$\hat{\boldsymbol{\mu}}^{t'}_{\mathbf{p}} = \boldsymbol{\mu}^{t'}_{\mathbf{p}} + \hat{\boldsymbol{\mu}}_{\Delta\mathbf{p}}, \quad \hat{\mathbf{S}}^{t'} = \mathbf{S} + \hat{\boldsymbol{\mu}}_{\Delta\mathbf{S}}, \quad \hat{\mathbf{R}}^{t'} = \mathbf{R}^{t'} \oplus \hat{\boldsymbol{\mu}}_{\Delta\mathbf{R}}, \tag{13}$$

where $\hat{\boldsymbol{\mu}}_{\Delta\mathbf{p}}, \hat{\boldsymbol{\mu}}_{\Delta\mathbf{S}}, \hat{\boldsymbol{\mu}}_{\Delta\mathbf{R}}$ denote the inferred variations in position, scale, and rotation, respectively, and $\oplus$ is quaternion-based rotation composition. In parallel, we modulate each primitive's opacity based on its proximity to the canonical state in both time and orientation. This confidence-weighted visibility is defined as:

$$\hat{\sigma}^{t'} = \sigma^{t'} \cdot \exp\left(-\frac{1}{2}\left[\Sigma_t^{-1}(t' - \mu_t)^2 + \left(\mathbf{O}^{t'} \ominus \boldsymbol{\mu}_{\mathbf{O}}\right)^{\top} \boldsymbol{\Sigma}_{\mathbf{O}}^{-1}\left(\mathbf{O}^{t'} \ominus \boldsymbol{\mu}_{\mathbf{O}}\right)\right]\right). \tag{14}$$

Together, these modulations encourage each hyper-Gaussian to adaptively express geometry and appearance consistent with the temporal and orientational cues, while suppressing noisy contributions from out-of-support regions.

Anchor-Guided Deformation and Conditioning. Recall that to query the hyper-Gaussian via conditioned slicing defined by Eq. (9), we must first estimate each Gaussian's local motion context, including its position and orientation under the evolving scene. We achieve this through anchor-driven deformation within our Global Orientation Field, drawing inspiration from Embedded Deformation Graphs [65, 69] and their extensions in dynamic modeling [5, 13, 44, 39, 52, 83].

Specifically, each Gaussian $\mathcal{G}_j$ is softly associated with a set of K nearby oriented anchors via skinning weights $\{w_{ij}\}$ computed from spatial proximity. Each anchor i carries a time-varying pair $(\mathbf{O}_i^t, \boldsymbol{\tau}_i^t)$, representing its orientation and position at time t. We express the relative anchor transformation in $SE(3)$ across time as: $\Delta\mathbf{Q}_i^{1\to t'} = (\mathbf{O}_i^{t'}, \boldsymbol{\tau}_i^{t'})(\mathbf{O}_i^1, \boldsymbol{\tau}_i^1)^{-1}$, which is further converted into a dual quaternion $\mathbf{q}_i^{1\to t'} \in \mathbb{D}$. To deform $\mathcal{G}_j$ towards any target time t', we aggregate the relative transformations of its neighboring anchors using a weighted dual quaternion blend [52]: $\hat{\mathbf{q}}_j^{1\to t'} = \text{Blend}_{\mathbb{D}}\left(\{w_{ij}, \mathbf{q}_i^{1\to t'}\}_{i=1}^K\right)$, which then yields the updated attributes:

$$(\mathbf{R}_j^{t'}, \boldsymbol{\mu}_{\mathbf{p}\,j}^{t'}) = \hat{\mathbf{q}}_j^{1\to t'}(\mathbf{R}_j^1, \boldsymbol{\mu}_{\mathbf{p}\,j}^1). \tag{15}$$

Importantly, this deformation scheme provides access to the local orientation $\mathbf{O}^{t'}$ required for state conditioning in Eq. (9). However, this requires additional care since the deformed Gaussian may not coincide with any specific anchor. Therefore, we interpolate $\mathbf{O}^{t'}$ from surrounding anchors:

$$\mathbf{O}^{t'} = \text{Blend}_{SO(3)}\left(\{w_{ij}, \mathbf{O}_i^{t'}\}_{i=1}^K\right). \tag{16}$$

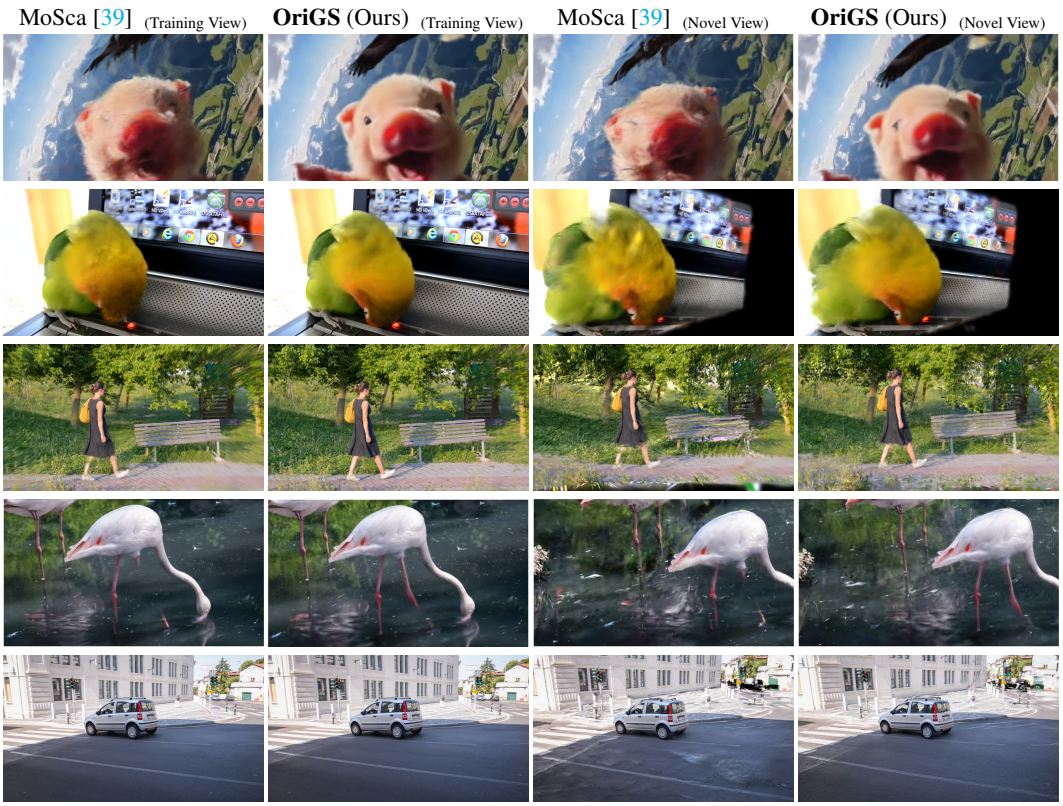

Figure 3: **Visual comparisons of training sequence reconstruction and novel view synthesis on in-the-wild videos.** Please 🔍 zoom in for more details.

This produces a smoothly varying orientation field that respects both the global structure and local dynamics. By grounding the conditioned slicing on these anchor-driven estimates, our hyper-Gaussian representation can adaptively capture complex motion across different regions.

5 Experiments

5.1 Experimental Protocol

Datasets. We evaluate the 4D reconstruction and novel view synthesis performance of our OriGS on a diverse set of in-the-wild monocular videos, emphasizing scenes with complex motion and object interactions. Our primary evaluation includes videos from DAVIS [59], OpenAI SORA [6], and YouTube-VOS [74], which reflect the casual and unconstrained nature of our target setting. For quantitative benchmarking, we additionally report results on the DyCheck dataset [19], which contains seven scenes with multi-camera captures for novel view synthesis evaluation.

Evaluation. For in-the-wild videos, where ground-truth novel views are inherently unavailable, we conduct qualitative evaluations to assess the visual fidelity of reconstructed scenes. For the DyCheck dataset [19], which provides multi-camera setups for quantitative assessment, we evaluate our method using standard image quality metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [68], and Learned Perceptual Image Patch Similarity (LPIPS) [78]. The DyCheck dataset uses three synchronized cameras: one hand-held mobile view and two stationary references. Following [19], we measure reconstruction quality from the reference viewpoints.

5.2 Experimental Results

In-the-wild. We qualitatively evaluate OriGS on a range of casually captured monocular videos that exhibit diverse motion patterns. Figure 3 presents side-by-side comparisons with MoSca [39], a recent state-of-the-art system for automatic 4D reconstruction in real-world scenarios. Across

Table 1: **Quantitative comparisons of novel view synthesis on DyCheck.** We report PSNR, SSIM, and LPIPS metrics across seven scenes. Markers ○ and ● denote with and without ground-truth camera pose, respectively. The best, second best, and third best results are highlighted.

Method	Apple			Block			Paper Windmill			Space Out		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
○ T-NeRF [19]	17.43	0.728	0.508	17.52	0.669	0.346	17.55	0.367	0.258	17.71	0.591	0.377
○ NSFF [42]	16.47	0.754	0.414	14.71	0.606	0.438	14.94	0.272	0.348	17.65	0.636	0.341
○ Nerfies [55]	17.54	0.750	0.478	16.61	0.639	0.389	17.34	0.378	0.211	17.79	0.622	0.303
○ HyperNeRF [56]	17.64	0.743	0.478	17.54	0.670	0.331	17.38	0.382	0.209	17.93	0.605	0.320
● RoDynRF [47]	18.73	0.722	0.552	18.73	0.634	0.513	16.71	0.321	0.482	18.56	0.594	0.413
○ DynPoint [80]	17.78	0.743	-	17.67	0.667	-	17.32	0.366	-	17.78	0.603	-
○ PGDVS [79]	16.66	0.721	0.411	16.38	0.601	0.293	17.19	0.386	0.277	16.49	0.592	0.326
○ DyBluRF [7]	18.00	0.737	0.488	17.47	0.665	0.349	18.19	0.405	0.301	18.83	0.643	0.326
○ CTNeRF [50]	19.53	0.691	-	19.74	0.626	-	17.66	0.346	0.346	18.11	0.601	-
○ Dyn. GS [49]	7.65	-	0.766	7.55	-	0.684	6.24	-	0.729	6.79	-	0.733
○ 4DGS [71]	15.41	-	0.450	11.28	-	0.633	15.60	-	0.297	14.60	-	0.372
○ D-NPC [35]	16.83	0.752	0.469	15.53	0.632	0.350	18.01	0.432	0.209	18.69	0.640	0.246
● Marbles [64]	16.50	-	0.499	16.11	-	0.363	16.19	-	0.454	15.97	-	0.437
○ SoM [67]	-	-	-	-	-	-	-	-	-	-	-	-
● MoSca [39]	15.99	0.705	0.503	18.20	0.665	0.324	21.37	0.645	0.175	22.57	0.750	0.198
● **OriGS** (Ours)	17.21	0.739	0.428	18.67	0.676	0.317	21.48	0.646	0.171	22.78	0.754	0.194
○ **OriGS** (Ours)	19.46	0.807	0.341	18.76	0.690	0.319	22.46	0.751	0.152	20.87	0.672	0.249

Method	Spin			Teddy			Wheel			**Average**		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
○ T-NeRF [19]	19.16	0.567	0.443	13.71	0.570	0.429	15.65	0.548	0.292	16.96	0.577	0.379
○ NSFF [42]	17.26	0.540	0.371	12.59	0.537	0.527	14.59	0.511	0.331	15.46	0.551	0.396
○ Nerfies [55]	18.38	0.585	0.309	13.65	0.557	0.372	13.82	0.458	0.310	16.45	0.570	0.339
○ HyperNeRF [56]	19.20	0.561	0.325	13.97	0.568	0.350	13.99	0.455	0.310	16.81	0.569	0.332
● RoDynRF [47]	17.41	0.484	0.570	14.33	0.536	0.613	15.20	0.449	0.478	17.10	0.534	0.517
○ DynPoint [80]	19.04	0.564	-	13.95	0.551	-	14.72	0.515	-	16.89	0.573	-
○ PGDVS [79]	18.49	0.590	0.247	13.29	0.516	0.399	12.68	0.429	0.429	15.88	0.548	0.340
○ DyBluRF [7]	18.20	0.541	0.400	14.61	0.572	0.435	16.26	0.575	0.325	17.37	0.591	0.373
○ CTNeRF [50]	19.79	0.516	-	14.51	0.509	-	14.48	0.430	-	17.69	0.531	-
○ Dyn. GS [49]	8.08	-	0.651	7.41	-	0.690	7.28	-	0.593	7.29	-	0.692
○ 4DGS [71]	14.42	-	0.339	12.36	-	0.466	11.79	-	0.436	13.64	-	0.428
○ D-NPC [35]	17.78	0.585	0.309	12.19	0.536	0.503	13.27	0.549	0.349	16.04	0.589	0.348
● Marbles [64]	17.51	-	0.424	13.68	-	0.443	14.58	-	0.389	15.79	-	0.428
○ SoM [67]	-	-	-	-	-	-	-	-	-	17.32	0.598	0.296
● MoSca [39]	20.71	0.677	0.229	15.54	0.624	0.356	17.74	0.667	0.238	18.84	0.676	0.289
● **OriGS** (Ours)	20.92	0.670	0.215	16.11	0.639	0.334	17.92	0.670	0.233	19.30	0.685	0.270
○ **OriGS** (Ours)	21.62	0.759	0.177	16.45	0.649	0.325	18.19	0.684	0.226	19.69	0.716	0.256

both training sequence reconstruction and novel view synthesis, OriGS consistently recovers cleaner geometry and preserves structural integrity under complex dynamics. Notably, our method remains robust in challenging scenarios, such as fast movement (*e.g.*, the spinning pig and the parrot). These results demonstrate that OriGS generalizes effectively to complex real-world videos, enabling faithful 4D reconstruction and compelling view synthesis without relying on multi-view constraints.

DyCheck. We compare OriGS against recent state-of-the-art methods, including NeRF-based [19, 42, 55, 56, 47, 7, 50, 80, 79] and 3DGS-based [49, 71, 35, 64, 67, 39]. Table 1 reports quantitative results across all seven scenes using PSNR, SSIM, and LPIPS. We indicate whether ground-truth camera poses are used by markers ○ (with) and ● (without). Our OriGS achieves superior reconstruction fidelity, especially in scenes involving rotational deformation (*e.g.*, the "Apple" and "Paper Windmill" scenes), where prior methods tend to produce artifacts. Notably, OriGS outperforms all baselines in the average across all three metrics, even without access to ground-truth camera pose. When provided with accurate poses, performance is further improved in most scenes, demonstrating the effectiveness of our framework. A minor exception is the "Space Out" scene, which is limited by inaccurate pose metadata provided in the dataset, as previously observed in [67].

Figure 4 showcases the qualitative advantages of OriGS in novel view synthesis. Compared to mainstream methods, our OriGS can produce sharper surface boundaries and more coherent geometry under challenging motion. For example, in the "Paper Windmill" scene (top row), other methods typically struggle with blurred or broken structures due to the rapid spinning motion. In contrast, our OriGS, by anchoring complex dynamics to a coherent orientation field, maintains the integrity of the rotating blades. Compared to OriGS, baseline results often exhibit severe ghosting artifacts and

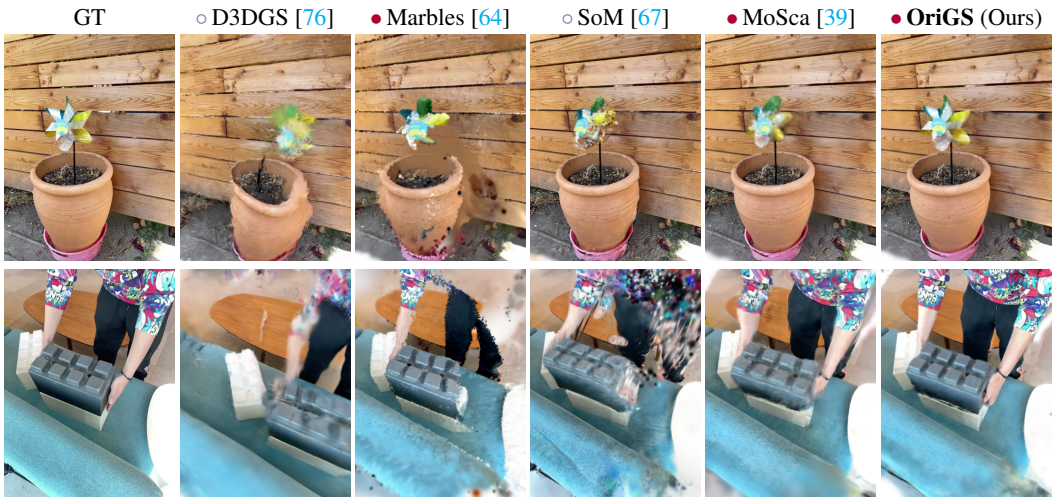

Figure 4: **Visual comparisons of novel view synthesis on DyCheck.** Markers ○ and ● denote with and without ground-truth camera pose, respectively. Our OriGS can recover sharper geometry and more coherent motion.

rigid-body distortion in scenes with hand-object occlusion (*e.g.*, the "Block" scene (bottom row)), further emphasizing the superiority of our method in handling complex motion.

5.3 Ablation Study

To analyze the impact of each design component in OriGS, we conduct a progressive ablation study from a minimal baseline to our full model. Table 2 reports quantitative results on the "Apple" scene from DyCheck [19] using the provided camera pose, and Figure 5 presents qualitative comparisons on the "Libby" scene from DAVIS [59]. We follow the same experimental setting for novel view synthesis as in Section 5.2. **(i)** We begin with **3DGS-MLP** as the baseline, where each 3D Gaussian primitive is updated individually across time using a shared deformation MLP, without explicit motion modeling. This baseline can only capture simple movement and struggles with structural coherence. **(ii)** Next, we consider **Deform w/ GOF**, which replaces the MLP with anchor-driven deformation guided by our Global Orientation Field. This variant enables Gaussian primitives to capture global motion, yet lacks the expressiveness to model complex dynamics across regions. **(iii)** To enable localized, time-adaptive behavior, we augment the model with **Hyper-Gaussian w/** t, where each Gaussian encodes a probabilistic state over time, space, and geometry (*i.e.*, $\boldsymbol{\xi} = (\Delta\mathbf{p}, \Delta\mathbf{g}, t)$). The deformation is inferred by conditioning this hyper-Gaussian on the target time step, enabling smoother and temporally adaptive reconstruction. **(iv)** Finally, our **OriGS** further incorporates orientation as a conditioning signal. By anchoring local dynamics to both temporal and orientational cues, it enables the model to capture direction-sensitive motion patterns and better preserve coherent evolution across regions with diverse dynamics.

Table 2: **Ablation study on OriGS variants on the "Apple" scene from DyCheck [19].**

Method	PSNR↑	SSIM↑	LPIPS↓
(i) 3DGS-MLP (Baseline)	13.47	0.532	0.586
(ii) Deform w/ GOF	16.28	0.694	0.476
(iii) Hyper-Gaussian w/ t	18.71	0.750	0.393
(iv) **OriGS** (Full)	**19.46**	**0.807**	**0.341**

Figure 5: **Ablation study on OriGS variants on the "Libby" scene from DAVIS [59].** We show a frame from the original input video (leftmost) and zoomed-in novel view synthesis results from different ablated variants. The main region of interest is the dog behind the pillar, which exhibits fast movement and heavy occlusion.

6 Conclusion

In this work, we presented Orientation-anchored Gaussian Splatting (OriGS), a novel framework for 4D reconstruction from casually captured monocular videos. Our core insight is to embrace orientation as a dynamic anchor that organizes how different parts of the scene deform and interact. By introducing a Global Orientation Field and a Hyper-Gaussian representation, OriGS models long-range evolution and local motion variation in a unified and principled manner, thereby capturing coherent dynamics across the scene. Extensive experiments demonstrate the superiority of our OriGS, enabling high-fidelity reconstruction and view synthesis under challenging real-world scenarios.

Acknowledgments: This research is supported by NSF IIS-2525840, CNS-2432534, ECCS-2514574, NIH 1RF1MH133764-01 and Cisco Research unrestricted gift. This article solely reflects opinions and conclusions of authors and not funding agencies.

References

[1] Marc Alexa, Daniel Cohen-Or, and David Levin. As-rigid-as-possible shape interpolation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 165–172. 2023.

[2] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16610–16620, 2023.

[3] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In *ECCV*, pages 321–335. Springer, 2024.

[4] Yanqi Bao, Tianyu Ding, Jing Huo, Yaoli Liu, Yuxin Li, Wenbin Li, Yang Gao, and Jiebo Luo. 3d gaussian splatting: Survey, technologies, challenges, and opportunities. In *IEEE TCSVT*, 2025.

[5] Aljaz Bozic, Pablo Palafox, Michael Zollhöfer, Angela Dai, Justus Thies, and Matthias Nießner. Neural non-rigid tracking. *Advances in Neural Information Processing Systems*, 33:18727–18737, 2020.

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

[7] Minh-Quan Viet Bui, Jongmin Park, Jihyong Oh, and Munchurl Kim. Dyblurf: Dynamic deblurring neural radiance fields for blurry monocular video. *arXiv preprint arXiv:2312.13528*, 2023.

[8] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023.

[9] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024.

[10] Jianmei Dai, Zhilong Zhang, Shiwen Mao, and Danpu Liu. A view synthesis-based 360° vr caching system over mec-enabled c-ran. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3843–3855, 2019.

[11] Stavros Diolatzis, Tobias Zirr, Alexander Kuznetsov, Georgios Kopanas, and Anton Kaplanyan. N-dimensional gaussians for fitting of high dimensional functions. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[12] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, et al. Bootstap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, pages 3257–3274, 2024.

[13] Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, and Shahram Izadi. 3d scanning deformable objects with a single rgbd sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 493–501, 2015.

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

[15] Bardienus P Duisterhof, Zhao Mandi, Yunchao Yao, Jia-Wei Liu, Mike Zheng Shou, Shuran Song, and Jeffrey Ichnowski. Md-splatting: Learning metric deformation from 4d gaussians in highly deformable scenes. *arXiv preprint arXiv:2312.00583*, 2(3), 2023.

[16] Denis J Evans. On the representation of orientation space. *Molecular physics*, 34(2):317–325, 1977.

[17] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022.

[18] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

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

[20] Kyle Gao, Yina Gao, Hongjie He, Dening Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022.

[21] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

[22] Zhongpai Gao, Benjamin Planche, Meng Zheng, Anwesa Choudhuri, Terrence Chen, and Ziyan Wu. 6dgs: Enhanced direction-aware gaussian splatting for volumetric rendering. In *ICLR*, 2025.

[23] Zhongpai Gao, Benjamin Planche, Meng Zheng, Anwesa Choudhuri, Terrence Chen, and Ziyan Wu. 7dgs: Unified spatial-temporal-angular gaussian splatting. *arXiv preprint arXiv:2503.07946*, 2025.

[24] Zhongpai Gao, Meng Zheng, Benjamin Planche, Anwesa Choudhuri, Terrence Chen, and Ziyan Wu. Render-fm: A foundation model for real-time photorealistic volumetric rendering. *arXiv preprint arXiv:2505.17338*, 2025.

[25] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):285–321, 1991.

[26] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[27] David Halliday, Robert Resnick, and Jearl Walker. *Fundamentals of physics*. John Wiley & Sons, 2013.

[28] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[29] 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.

[30] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023.

[31] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4220–4230, 2024.

[32] Takeo Igarashi, Tomer Moscovich, and John F Hughes. As-rigid-as-possible shape manipulation. *ACM transactions on Graphics (TOG)*, 24(3):1134–1141, 2005.

[33] Anastasia Ioannidou, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Deep learning advances in computer vision with 3d data: A survey. *ACM computing surveys (CSUR)*, 50(2):1–38, 2017.

[34] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, pages 402–418. Springer, 2022.

[35] Moritz Kappel, Florian Hahlbohm, Timon Scholz, Susana Castillo, Christian Theobalt, Martin Eisemann, Vladislav Golyanik, and Marcus Magnor. D-npc: Dynamic neural point clouds for non-rigid view synthesis from monocular video. In *Computer Graphics Forum*, page e70038. Wiley Online Library, 2024.

[36] 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.

[37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM Trans. Graph.*, 2023.

[38] Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic gaussian splatting. *NeurIPS*, 37:5384–5409, 2024.

[39] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *CVPR*, 2025.

[40] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5521–5531, 2022.

[41] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8508–8520, 2024.

[42] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. 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*, pages 6498–6508, 2021.

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

[44] Yiming Liang, Tianhan Xu, and Yuta Kikuchi. Himor: Monocular deformable gaussian reconstruction with hierarchical motion representation. In *CVPR*, 2025.

[45] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5741–5751, 2021.

[46] Qingming Liu, Yuan Liu, Jiepeng Wang, Xianqiang Lyv, Peng Wang, Wenping Wang, and Junhui Hou. Modgs: Dynamic gaussian splatting from casually-captured monocular videos. In *ICLR*, 2025.

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023.

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

[49] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024.

[50] Xingyu Miao, Yang Bai, Haoran Duan, Fan Wan, Yawen Huang, Yang Long, and Yefeng Zheng. Ctnerf: Cross-time transformer for dynamic neural radiance field from monocular video. *Pattern Recognition*, 156:110729, 2024.

[51] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[52] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015.

[53] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[54] Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. In *CVPR*, 2025.

[55] 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.

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

[57] Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, and Nahyup Kang. Temporal interpolation is all you need for dynamic neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4212–4221, 2023.

[58] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024.

[59] 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.

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

[61] Ahmed A Shabana. *Computational dynamics*. John Wiley & Sons, 2009.

[62] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023.

[63] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116. Citeseer, 2007.

[64] 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. In *SIGGRAPH Asia*, 2024.

[65] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es, 2007.

[66] Fengrui Tian, Shaoyi Du, and Yueqi Duan. Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17903–17913, 2023.

[67] 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.

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[69] Ofir Weber, Olga Sorkine, Yaron Lipman, and Craig Gotsman. Context-aware skeletal shape deformation. In *Computer Graphics Forum*, pages 265–274. Wiley Online Library, 2007.

[70] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 16210–16220, 2022.

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

[72] Zirui Wu, Tianyu Liu, Liyi Luo, Zhide Zhong, Jianteng Chen, Hongmin Xiao, Chao Hou, Haozhe Lou, Yuantao Chen, Runyi Yang, et al. Mars: An instance-aware, modular and realistic simulator for autonomous driving. In *CAAI International Conference on Artificial Intelligence*, pages 3–15. Springer, 2023.

[73] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20417, 2024.

[74] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018.

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

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

[77] Jihwan Yoon, Sangbeom Han, Jaeseok Oh, and Minsik Lee. Splinegs: Learning smooth trajectories in gaussian splatting for dynamic scene reconstruction. In *ICLR*, 2025.

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

[79] Xiaoming Zhao, Alex Colburn, Fangchang Ma, Miguel Angel Bautista, Joshua M Susskind, and Alexander G Schwing. Pseudo-generalized dynamic view synthesis from a video. *arXiv preprint arXiv:2310.08587*, 2023.

[80] Kaichen Zhou, Jia-Xing Zhong, Sangyun Shin, Kai Lu, Yiyuan Yang, Andrew Markham, and Niki Trigoni. Dynpoint: Dynamic neural point for view synthesis. *Advances in Neural Information Processing Systems*, 36:69532–69545, 2023.

[81] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[82] Jiaxuan Zhu and Hao Tang. Dynamic scene reconstruction: Recent advance in real-time rendering and streaming. *arXiv preprint arXiv:2503.08166*, 2025.

[83] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33(4):1–12, 2014.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction accurately reflect the contributions of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to Section 5.1 and supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code will be released publicly.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in this paper adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to supplementary materials.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All external assets such as datasets and code libraries used in the research are properly credited, and their licenses are respected and clearly mentioned.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Evaluation on Point Tracking

Following the DyCheck benchmark [19] and prior works [64, 39], we perform a quantitative evaluation on point tracking accuracy. We compare with both tracking and reconstruction methods. Results shown in Table 3 indicate the superiority of our OriGS.

Table 3: Correspondence Evaluation on DyCheck with PCK-T @0.05%.

Method	Nerfies [55]	HyperNeRF [56]	Dyn. GS [49]	4DGS [71]
PCK-T ↑	0.400	0.453	0.079	0.073
Method	CoTracker [36]	Marbles [64]	MoSca [39]	**OriGS** (Ours)
PCK-T ↑	0.803	0.806	0.824	**0.851**

B Demo Videos

To further showcase the effectiveness of OriGS in real-world scenarios, we include additional video demonstrations of in-the-wild 4D reconstruction in the supplementary material. Following the same experimental setup as in the main paper, we present visualizations of both training sequence reconstruction and novel view synthesis, comparing OriGS with the recent state-of-the-art 4D reconstruction system. In addition, we provide video demonstrations of the OriGS variants designed in our ablation study, illustrating how individual components affect reconstruction quality. These results collectively highlight the temporal consistency and structural fidelity of OriGS across diverse scenes and motion patterns.

C Implementation Details

Oriented Anchor Initialization. We follow a similar initialization pipeline provided in recent open-source codebases of dynamic reconstruction frameworks [64, 39]. Specifically, we first extract long-range 2D point trajectories using SpatialTracker [73] and obtain per-frame depth maps using DepthCrafter [29]. These 2D correspondences are lifted into 3D space using estimated camera parameters from bundle adjustment [39]. The resulting 3D trajectories are then converted into oriented anchors, whose initial principal directions are estimated by applying PCA over early-frame motion. In our experiments, we set the temporal window size $W = 5$. This step provides a stable estimate of forward direction, which is then temporally propagated to form Global Orientation Field.

Hyper-Gaussian Parameterization. We adopt the 3D Gaussian Splatting (3DGS) [37] as the base representation and extend each Gaussian primitive with a hyper-Gaussian parameterized by a canonical state $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and a Cholesky-decomposed covariance $\boldsymbol{\Sigma}_{\boldsymbol{\xi}} = \mathbf{L}_{\boldsymbol{\xi}}\mathbf{L}_{\boldsymbol{\xi}}^{\top}$, where $\mathbf{L}_{\boldsymbol{\xi}}$ is a lower-triangular matrix [22, 23, 11, 24]. To improve computational efficiency, we avoid constructing the full covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\xi}}$. Instead, we further factor it into two learnable subcomponents: **(i)** a marginal covariance $\boldsymbol{\Sigma}_{(t,\mathbf{O})}$ over temporal and orientational variables, which inherits the Cholesky-decomposed parameterization to ensure positive semi-definiteness, and **(ii)** a cross-covariance term $\boldsymbol{\Sigma}_{(\Delta\mathbf{p},\Delta\mathbf{g}),(t,\mathbf{O})}$ that captures the correlation between dynamic geometry $(\Delta\mathbf{p}, \Delta\mathbf{g})$ and temporal-orientational context. During training, we jointly optimize these covariance terms along with the canonical mean $\boldsymbol{\mu}_{\boldsymbol{\xi}}$ and the original 3DGS parameters (position, scale, rotation, color, and opacity). This factorization strategy circumvents the overhead of full matrix parameterization while retaining the necessary statistical coupling for efficient conditioned slicing.

Optimization. We optimize the model using the differentiable rasterization-based rendering pipeline from 3DGS [37], adapted to support high-dimensional slicing and dynamic modulation. The loss function combines: **(i)** photometric loss [37], an RGB reconstruction loss between rendered and ground-truth images, **(ii)** 2D correspondence loss, alignment to long-range 2D tracks and depth priors from foundation models, as in [64, 39, 44, 54], and **(iii)** deformation regularization, an as-rigid-as-possible constraint [73, 63, 1, 32] applied to anchor-guided transformations. For scalability and high-fidelity reconstruction, we also design a pruning-and-densification scheme: Gaussian primitives with low opacity are pruned, while spatial regions exhibiting high response gradients with respect to $\boldsymbol{\mu}_{\mathbf{p}}$ and $\boldsymbol{\mu}_{\Delta\mathbf{p}}$ are densified through local duplication. All experiments are conducted on a single

NVIDIA RTX A6000 GPU, and the full optimization of a typical scene takes approximately 0.5–2 hours, depending on video length and complexity.

D Limitations

While OriGS demonstrates promising results, it also presents certain limitations. **(i)** In scenes dominated by simple motion, such as near-rigid translation along a fixed axis, orientation tends to remain approximately constant over time. Consequently, conditioning on these orientational cues may offer limited additional benefit, yet our framework still incurs unnecessary optimization complexity of high-dimensional modeling. Future work could explore adaptive mechanisms that modulate the use or dimensionality of orientation cues based on the complexity of motion in the scene. **(ii)** OriGS leverages 2D priors such as point trajectories and depth from vision foundation models to initialize and optimize the orientation field. The reliability of these priors can affect the quality of 4D reconstruction. This reflects a deeper challenge tied to the development of reliable visual priors, which has long been a cornerstone of progress in computer vision research.

E Broader Impacts

OriGS provides a unified framework for reconstructing dynamic scenes from casual monocular videos, which can benefit various real-world applications. For instance, in virtual and augmented reality, OriGS can reconstruct dynamic environments or actors from consumer-grade video input alone, lowering the barrier to immersive content creation. Our work can also support robotics and behavioral analysis, especially where low-cost monocular capture is the only viable option. However, as with other scene reconstruction techniques, OriGS could potentially be misused for unauthorized replication of environments or individuals. While our framework is not designed for such misconduct, responsible usage should consider privacy and ethical concerns.