DRoPS: Dynamic 3D Reconstruction of Pre-Scanned Objects

Narek Tumanyan[1,2], Samuel Rota Bulò[2], Denis Rozumny[2], Lorenzo Porzi[2], Adam Harley[2], Tali Dekel[1], Peter Kontschieder[2], and Jonathon Luiten[2]

[1] Weizmann Institute of Science
[2] Meta Reality Labs
Project webpage: drops-dynamics.github.io

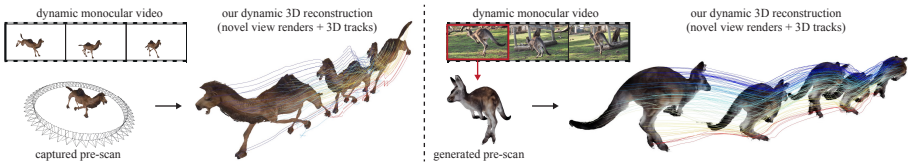

Fig. 1: Given a static pre-scan and a monocular video of a dynamic object, DRoPS reconstructs a complete dynamic 3D representation, enabling high-quality novel view synthesis. The pre-scan is either captured (left) or is generated from the first monocular video frame (right). The figure shows the dynamic objects rendered at multiple timesteps from a fixed novel view. The colored lines visualize the 3D point trajectories that emerge from our dynamic model.

Abstract. Dynamic scene reconstruction from casual videos has seen recent remarkable progress. Numerous approaches have attempted to overcome the ill-posedness of the task by distilling priors from 2D foundational models and by imposing hand-crafted regularization on the optimized motion. However, these methods struggle to reconstruct scenes from extreme novel viewpoints, especially when highly articulated motions are present. In this paper, we present *DRoPS* – a novel approach that leverages a static pre-scan of the dynamic object as an explicit geometric and appearance prior. While existing state-of-the-art methods fail to fully exploit the pre-scan, DRoPS leverages our novel setup to effectively constrain the solution space and ensure geometrical consistency throughout the sequence. The core of our novelty is twofold: first, we establish a grid-structured and surface-aligned model by organizing Gaussian primitives into pixel grids anchored to the object surface. Second, by leveraging the grid structure of our primitives, we parameterize motion using a CNN conditioned on those grids, injecting strong implicit regularization and correlating the motion of nearby points. Extensive experiments demonstrate that our method significantly outperforms the current state of the art in rendering quality and 3D tracking accuracy.

1 Introduction

Reconstruction of dynamic 3D subjects from casual video is transformative for many industries, from unlocking new forms of 3D content to metric-space understanding of how objects move, with applications in robotics, augmented reality,

and scene analysis. While significant progress has been made, existing approaches often struggle to maintain high-fidelity reconstruction and consistent novel view synthesis in casual video capture settings, particularly when rendered from extreme viewpoints far from the input camera trajectory.

Existing methods for dynamic reconstruction span a spectrum from fully multi-view setups to purely monocular approaches. Multi-view methods [21, 31, 41, 65] can achieve excellent results but require expensive synchronized camera arrays that are impractical for casual capture. Monocular methods [12, 27, 45, 59] are highly practical but face severe ill-posedness: infinitely many 3D motions can explain the same 2D observations. In this paper, we propose a novel setting that bridges this gap: monocular dynamic capture combined with a static pre-scan of the object. The pre-scan is either provided as part of the monocular sequence or generated by an image-to-3D model—a field that continues to advance rapidly. This setting is nearly as practical as monocular, while providing crucial geometric constraints that dramatically reduce ambiguity.

In this paper, we present such an approach by building upon 3D Gaussian Splatting [23] (3DGS), which has recently emerged as a powerful representation for novel view synthesis. 3DGS represents scenes as collections of colored 3D Gaussians that are rendered via differentiable splatting, enabling real-time rendering and efficient optimization. Recent works have extended 3DGS to dynamic scenes [41, 65, 70], but these typically require multi-view input or struggle with geometric accuracy from novel viewpoints. We extend this paradigm to the novel pre-scan-plus-monocular setting. Our approach consists of two main components: (i) constructing structured, surface-aligned Gaussians from the pre-scan by organizing them on pixel grids anchored to the object surface, and (ii) modeling the deformation of Gaussians through a convolutional network (CNN).

A key insight of our approach is that by organizing canonical Gaussians into structured, surface-aligned pixel grids and by supervising them with lifted 2D tracks, we ensure that each primitive persistently represents the same surface point throughout the sequence. This contrasts with prior 3DGS representations, which typically consist of unordered primitives and/or lack surface representation. While works like PixelSplat [3] organize Gaussians on pixel grids derived from input images, they do not anchor these primitives to a consistent object manifold. Conversely, while methods such as SuGaR [13] and DM4D [34] encourage Gaussians to align with the object surface, their underlying representation remains unorganized. Our persistent, surface-aligned, and grid-structured modeling is critical for achieving geometrically faithful reconstructions and high-quality novel view synthesis, particularly from viewpoints not observed in the input video.

Our grid-structured Gaussians are constructed on virtual camera planes oriented toward the canonical pre-scan 3DGS representation. By estimating appearance and depth from these virtual viewpoints, we back-project each pixel to a Gaussian, yielding a set of surface-aligned primitives organized on pixel grids.

Building on this structured representation, we introduce a Deep Motion Prior (DMP), a CNN that maps the canonical Gaussian grids to per-timestep

6-DOF deformations. By operating on structured pixel grids rather than unordered point sets, our approach leverages the strong spatial inductive bias of Convolutional networks: their translation equivariance and local connectivity naturally enforce local smoothness, correlating the motion of adjacent surface points and favoring spatially coherent motion fields, without heavily relying on explicit handcrafted regularization. Consequently, as demonstrated in Fig. 1, our design enables geometrically consistent dynamic reconstructions of highly articulated scenes, achieving high-quality synthesis even from extreme novel views.

We evaluate DRoPS on real-world and synthetic captures, and demonstrate that it significantly outperforms prior state-of-the-art monocular dynamic reconstruction methods, including optimization and generative-based approaches, in photometric, perceptual, semantic consistency, and long-range 3D tracking.

To summarize, our core contributions are: (**1**) We introduce a new task of dynamic 3D reconstruction from casual videos with pre-scanned objects – a practical setting with significantly reduced ill-posedness. (**2**) We present a novel canonical model using surface-aligned 3D Gaussians organized into pixel grids, ensuring persistent representation of each surface point. (**3**) We propose a CNN-based motion parameterization (DMP) that leverages the convolutional inductive bias for implicit regularization. (**4**) On real-world and synthetic benchmarks, we significantly outperform prior state-of-the-art methods by over 1 dB in PSNR with substantial improvements in perceptual (LPIPS), semantic (CLIP), and 3D tracking metrics; systematic ablations confirm that each core design component contributes meaningfully, with removing any one of them causing a noticeable performance drop.

2 Related Work

Dynamic Novel-View Synthesis. The field of novel-view synthesis has seen remarkable progress over the years [40], especially after the introduction of Neural Radiance Fields (NeRF) [1, 30, 36, 43, 57]. Previous works fall into several categories: (**a**) Per-timestep methods [67] that fit independent representations for each frame without modeling temporal correspondence. (**b**) Eulerian approaches [2,10] that represent scenes on 4D space-time grids, enabling smooth interpolation but lacking explicit correspondence. (**c**) Canonical-plus-deformation methods [45–47] that learn a canonical representation and per-frame deformation fields. (**d**) Template-guided methods [20, 29, 62, 69] that leverage human body models or other priors. Related monocular approaches incorporate motion via scene flow or other temporal constraints to enable view and time synthesis from a single posed video [11, 12, 30, 32, 33]. Most related to our work are point-based Lagrangian methods that track individual primitives through time [41]. Our work falls in between the full multi-camera and full monocular settings by taking a static pre-scan and a monocular dynamic sequence as input. In contrast to these approaches, we initialize canonical, grid-structured, and surface-aligned Gaussians and employ a CNN-based motion prior for regularization.

Dynamic 3D Gaussians. 3D Gaussian Splatting [23] has emerged as a powerful representation for real-time novel view synthesis. Several works have extended this to dynamic scenes: Dynamic 3D Gaussians [41] tracks persistent Gaussians through time but requires multi-view input; 4D Gaussian Splatting [65] and Deformable 3D Gaussians [70] learn temporal deformations but struggle with geometric consistency from extreme viewpoints; SC-GS [19] introduces sparse control points for editable dynamic scenes. More recently, Shape-of-Motion [59] and MoSca [27] achieved SOTA results by leveraging a low-rank motion basis and incorporating foundation model priors for depth and 2D tracking. Their successors, HiMoR [35] and OriGS [66], introduced improvements by hierarchical and orientation-aware motion modeling. Similarly, we incorporate priors from 2D tracking and depth foundation models. Unlike these methods, we leverage a pre-scan prior and surface-aligned pixel-grid organization to ensure geometric consistency even from viewpoints far from the input camera. We show that SOTA dynamic 3DGS methods fail to effectively leverage the pre-scan information provided within the monocular sequence as an initial static scan preceding the dynamic motion. Moreover, instead of heavily relying on handcrafted motion priors, we leverage the strong implicit regularization of CNNs by parameterizing motion with a U-Net [50].

Deep Image Prior. [55] demonstrated that the architecture of CNNs provides a strong implicit prior for image restoration, even without any training data. This deep prior has been extended to various domains including image denoising [16], Bayesian inference [5], 3D reconstruction [42,63], and point tracking [54]. The key insight is that CNN architectures inherently favor smooth, natural signals due to their local connectivity and translation equivariance. We apply this principle to motion estimation: by parameterizing scene motion through a CNN operating on structured canonical parameter grids, we leverage the network's spatial inductive bias to regularize the ill-posed monocular motion estimation problem.

Point Tracking. Dense point tracking has advanced significantly with learning-based methods. PIPs [14] introduced a transformer-like approach for tracking points through occlusions; TAP-Vid [7] established benchmarks; TAPIR [8], CoTracker [22], and AllTracker [15] improved accuracy and efficiency. OmniMotion [58] and DINO-Tracker [54] are most related to our work, using test-time optimization to estimate dense motion fields, but these methods focus on 2D tracking and require optical flow or video input. In contrast, we leverage 2D tracking predictions as supervision for our 3D motion estimation, and our surface-aligned Gaussians provide 3D correspondence throughout the sequence.

3 Method

Given a monocular video of T frames $\mathcal{I}_{\text{video}} \coloneqq \{(\mathbf{I}_{\mathtt{V}}^t, \mathbf{D}_{\mathtt{V}}^t, \mathbf{T}_{\mathtt{V}}^t)\}_{t=1}^{T}$ of a dynamic object, where each timestep includes an RGB frame $\mathbf{I}_{\mathtt{V}}^t$, an estimated depth map $\mathbf{D}_{\mathtt{V}}^t$, and an estimated camera pose $\mathbf{T}_{\mathtt{V}}^t$, together with a pre-scan of R frames $\mathcal{I}_{\text{scan}} \coloneqq \{(\mathbf{I}_{\mathtt{S}}^r, \mathbf{D}_{\mathtt{S}}^r, \mathbf{T}_{\mathtt{S}}^r)\}_{r=1}^{R}$ of the same object in a static state (either captured by the monocular camera or generated by an image-to-3D model), our goal is

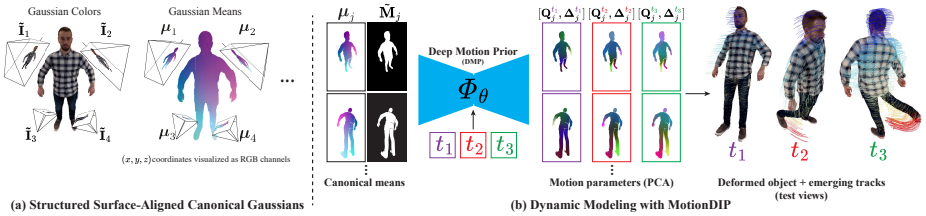

Fig. 2: *DRoPS Overview.* (a) We organize our canonical Gaussians into structured pixel grids that reside on virtual cameras surrounding the object, each pixel encoding the parameters of its back-projected 3D Gaussian (Sec. 3.1). (b) To reconstruct the dynamic sequence, we model the object deformation with Deep Motion Prior Φ_θ – a CNN that maps canonical positions $\boldsymbol{\mu}_j$ and timestep encodings $\gamma(t)$ to 6-DOF motion parameters $[\mathbf{Q}_j, \boldsymbol{\Delta}_j]$ (see Sec. 3.2). Timesteps t_1, t_2, t_3 are color-coded in the figure.

to obtain a complete dynamic 3D reconstruction enabling 360°-orbiting novel view synthesis. Depth maps and camera poses are obtained from off-the-shelf methods (see Sec. 4.1). As illustrated in Fig. 2, our method comprises two core components: (1) constructing canonical, surface-aligned Gaussians organized on pixel grids, where each primitive persistently represents a fixed surface point, and (2) modeling scene dynamics via a Deep Motion Prior (DMP)—a CNN that predicts per-timestep 6-DOF deformations for each Gaussian.

3.1 Pre-scan Canonical Model

Our canonical model leverages the static pre-scan to construct a structured 3D Gaussian representation that serves as the geometric foundation for dynamic reconstruction. We first build surface-aligned Gaussians organized into pixel grids, then refine them using photometric supervision. The pre-scan is either captured as part of the monocular sequence or, for in-the-wild fully monocular videos, generated from the first frame using an image-to-3D model [68]. Since image-to-3D generations are not inherently aligned with the video camera, we perform a pose alignment step using MASt3R [28] correspondences with Perspective-n-Points (PnP) to register the generated mesh to the first frame (see Appendix A.2 for details).

Structured Surface-Aligned Canonical Gaussians We use 3D Gaussian Splatting [23] to represent the canonical scene. Given the pre-scan $\mathcal{I}_{\text{scan}}$, we compute segmentation masks $\mathcal{M} \coloneqq \{\mathbf{M}_{\text{s}}^r\}_{r=1}^R$ using SAM2 [49]. We then fit a 3DGS representation $\mathcal{G}_{\text{init}} \coloneqq \{g_i \coloneqq (\mu_i, s_i, q_i, \alpha_i, c_i, m_i)\}_{i=1}^N$ with N Gaussians. Each Gaussian g_i is parameterized by its mean $\mu_i \in \mathbb{R}^3$, scale $s_i \in \mathbb{R}_+^3$, rotation quaternion $q_i \in \mathbb{R}^4$, opacity $\alpha_i \in [0, 1]$, and spherical harmonic coefficients c_i for view-dependent color. Additionally, $m_i \in [0, 1]$ is a foreground (FG) probability distilled from $\mathcal{M}$, used to separate dynamic from static Gaussians. We employ existing techniques for mask distillation [52, 72], which additionally reconstruct per-Gaussian mask channels during 3DGS optimization.

While standard Gaussian Splatting achieves high-fidelity photometric reconstruction, it lacks an explicit surface representation [6,17] and does not guarantee that primitives persistently correspond to the same surface points over time. To

address this, as illustrated in Fig. 2(a) and Fig. 3, we organize Gaussians into structured, surface-aligned pixel grids that anchor each primitive to a consistent surface location. We show that this persistent modeling is crucial for faithful and consistent dynamic reconstruction.

To obtain the surface-aligned representation from $\mathcal{G}_{\text{init}}$, we first compute a 3D bounding box $\mathcal{B}$ around the FG Gaussians $\mathcal{G}_{\text{fg}} = \{g_i \in \mathcal{G}_{\text{init}} : m_i > 0.5\}$. For each face j of $\mathcal{B}$, we define a virtual stereo camera pair looking inward toward the object center and render $\mathcal{G}_{\text{fg}}$ from both views. The resulting stereo image pairs are processed by FoundationStereo [61, 64] to estimate dense depth maps. We denote the rendered image, mask, and estimated depth from the left view as $\tilde{\mathbf{I}}_j$, $\tilde{\mathbf{M}}_j$, and $\tilde{\mathbf{D}}_j$, respectively, and extract DINOv2 features $\tilde{\mathbf{F}}_j$ [44] from $\tilde{\mathbf{I}}_j$. These depth maps define the object surface from each viewpoint. For each face j and each FG pixel (x, y) in $\tilde{\mathbf{I}}_j$, we create an isotropic Gaussian $g_{j,x,y}$ with color $c_{j,x,y} \coloneqq \tilde{\mathbf{I}}_j[x, y]$, DINOv2 feature $f_{j,x,y} \coloneqq \tilde{\mathbf{F}}_j[x, y]$, identity rotation $q_{j,x,y} \coloneqq [1, 0, 0, 0]$, fixed opacity $\alpha_{j,x,y} \coloneqq 0.98$, and position/scale derived from the estimated depth $\tilde{\mathbf{D}}_j[x, y]$:

$$\mu_{j,x,y} \coloneqq \tilde{\mathbf{T}}_j^{-1}(x, y)\tilde{\mathbf{D}}_j[x, y]\,, \qquad s_{j,x,y} \coloneqq \frac{\tilde{\mathbf{D}}_j[x, y]}{f_\ell} \cdot 0.95\,, \tag{1}$$

where $\tilde{\mathbf{T}}_j^{-1}(x, y) \in \mathbb{R}^3$ denotes pixel (x, y) unprojected to unit depth (i.e., $z = 1$) in camera space, and f_ℓ is the focal length. We denote the resulting set of surface-aligned Gaussians as $\mathcal{G}_{\text{can}} \coloneqq \{g_{j,x,y}\}$.

Canonical Refinement The depth estimates from FoundationStereo provide a good initialization for $\mathcal{G}_{\text{can}}$, but lack high-frequency surface details (see Appendix A.7 for visualizations). To recover fine-grained geometry, we refine the depth maps $\tilde{\mathbf{D}}_j$ and colors $\tilde{\mathbf{I}}_j$ that parameterize the canonical Gaussians by minimizing a photometric loss on the pre-scan views $\mathcal{I}_{\text{scan}}$. For each pre-scan view r, we define:

$$\mathcal{L}_{\text{can}}^r \coloneqq \mathcal{L}_{\text{L1}}(\hat{\mathbf{I}}_{\text{S}}^r, \mathbf{I}_{\text{S}}^r) + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}}(\hat{\mathbf{I}}_{\text{S}}^r, \mathbf{I}_{\text{S}}^r) + \lambda_{\text{TV}} \sum_j \mathcal{L}_{\text{TV}}(\tilde{\mathbf{I}}_j)\,, \tag{2}$$

where $\mathbf{I}_{\text{S}}^r$, $\hat{\mathbf{I}}_{\text{S}}^r$ are the ground-truth and rendered images from pre-scan camera $\mathbf{T}_{\text{S}}^r$, respectively. $\mathcal{L}_{\text{TV}}$ is a total variation regularizer that encourages spatial smoothness in the parameter grids. λ_* are trade-off parameters (values in Appendix A.4). The full canonical refinement loss averages over all pre-scan views: $\mathcal{L}_{\text{can}} \coloneqq \frac{1}{R}\sum_{r=1}^{R} \mathcal{L}_{\text{can}}^r$.

3.2 Deep Motion Prior

Estimating 3D motion from a monocular video is inherently ill-posed: many deformations satisfy the photometric loss while significantly distorting geometry. Existing methods address this with handcrafted priors such as as-rigid-as-possible regularization [19] or low-rank motion [59]. Instead, we harness the deep implicit prior of CNNs by leveraging the pixel-grid organization of $\mathcal{G}_{\text{can}}$.

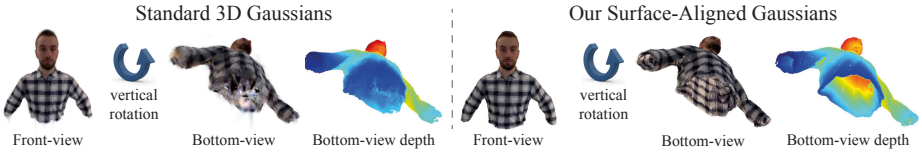

Fig. 3: *Surface-aligned Gaussians.* We visualize canonical Gaussians of the upper body. Unlike standard 3DGS (left), where Gaussians are unordered and lack surface representation, our structured Gaussians align with the object's surface, providing a more robust and generalizable representation for dynamic-time reconstruction.

We design a Deep Motion Prior (DMP) Φ_θ that takes as input the canonical position grids $\boldsymbol{\mu}_j \in \mathbb{R}^{H\times W\times 3}$, where $\boldsymbol{\mu}_j[x,y] \coloneqq \mu_{j,x,y}$, and outputs per-pixel 6-DOF deformations. For timestep t:

$$(\mathbf{Q}_j^t, \boldsymbol{\Delta}_j^t) \coloneqq \Phi_\theta(\boldsymbol{\mu}_j, t)\,, \tag{3}$$

where $\mathbf{Q}_j^t \in \mathbb{R}^{H\times W\times 4}$ and $\boldsymbol{\Delta}_j^t \in \mathbb{R}^{H\times W\times 3}$ are per-Gaussian rotation quaternions and translations, respectively. The deformed positions are computed as:

$$\boldsymbol{\mu}_j^t \coloneqq \mathrm{rot}(\boldsymbol{\mu}_j, \mathbf{Q}_j^t) + \boldsymbol{\Delta}_j^t\,, \tag{4}$$

where $\mathrm{rot}(\cdot,\cdot)$ applies the quaternion rotation pixel-wise. To prevent content duplication across parameter grids, we apply de-duplication masks on output grids $\boldsymbol{\mu}_j^t$ (see Appendix A.3 for details). The network Φ_θ is initialized to predict identity transformations, ensuring optimization starts from the canonical pose. The CNN's spatial inductive bias naturally enforces smooth, locally coherent deformations.

3.3 Optimization

This section enumerates the loss terms used to optimize our Deep Motion Prior.

Photometric Loss. For each timestep t, we render the deformed Gaussians from the input camera $\mathbf{T}_\mathtt{V}^t$, producing an image $\hat{\mathbf{I}}_\mathtt{V}^t$. We minimize the difference between the rendered and observed images using the standard 3DGS loss [23]:

$$\mathcal{L}_{\text{photo}}^t \coloneqq \mathcal{L}_{\text{L1}}(\hat{\mathbf{I}}_\mathtt{V}^t, \mathbf{I}_\mathtt{V}^t) + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}}(\hat{\mathbf{I}}_\mathtt{V}^t, \mathbf{I}_\mathtt{V}^t)\,, \tag{5}$$

where $\mathbf{I}_\mathtt{V}^t$ is the ground-truth frame from $\mathcal{I}_{\text{video}}$.

Point Tracking Loss. We leverage the prior of a pre-trained 2D point tracker (AllTracker [15]) to supervise the 2D foreground motion predicted by DMP. Given K foreground query points $\{p_k^1\}_{k=1}^K$ sampled at the reference frame ($t=1$), AllTracker provides ground-truth 2D correspondences p_k^t for each timestep t. Our model's prediction $\hat{p}_k^t$ is obtained by rasterizing the 2D positions of dynamic Gaussians at time t onto the reference frame [59]. The loss is defined as:

$$\mathcal{L}_{\text{track}}^t \coloneqq \frac{1}{K}\sum_{k=1}^{K} \|p_k^t - \hat{p}_k^t\|_1\,. \tag{6}$$

Depth Loss. For each timestep t, we render the depth of the deformed Gaussians from camera $\mathbf{T}_{\mathtt{V}}^t$, producing a depth map $\hat{\mathbf{D}}_{\mathtt{V}}^t$. Let $\mathbf{M}_{\mathtt{FG}}^t$ denote the intersection of the ground-truth and rendered foreground masks. We supervise the rendered depth against the input depth $\mathbf{D}_{\mathtt{V}}^t$ within this region:

$$\mathcal{L}_{\text{depth}}^t \coloneqq \frac{1}{\|\mathbf{M}_{\mathtt{FG}}^t\|_1} \|(\hat{\mathbf{D}}_{\mathtt{V}}^t - \mathbf{D}_{\mathtt{V}}^t) \odot \mathbf{M}_{\mathtt{FG}}^t\|_1 \,, \tag{7}$$

where $\odot$ denotes element-wise multiplication.

Depth Reprojection Loss. We additionally supervise the *reprojected* depth of dynamic points. Given a foreground query point p_k^1 in the reference frame ($t = 1$) and its tracked correspondence p_k^t in frame t, we obtain the ground-truth depth by sampling the input depth map: $\mathbf{D}_{\mathtt{V}}^t[p_k^t]$. To obtain our model's prediction, we render for each Gaussian its depth at time t but at the 2D location it occupies in frame 1, yielding the reprojected depth map $\hat{\mathbf{D}}^{t\to 1}$. The loss encourages consistency between tracked correspondences and predicted motion:

$$\mathcal{L}_{\text{reproj}}^t \coloneqq \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{D}_{\mathtt{V}}^t[p_k^t] - \hat{\mathbf{D}}^{t\to 1}[p_k^1]\|_1 \,. \tag{8}$$

Isometry Losses. To preserve geometric consistency during deformation, we introduce multi-scale isometry losses that encourage distance preservation between Gaussian pairs at different spatial scales. For notational clarity, we re-index the canonical Gaussians $g_{j,x,y} \in \mathcal{G}_{\text{can}}$ with a single index $\mathtt{i}$, writing $g_{\mathtt{i}}$ to denote an arbitrary Gaussian, $\mu_{\mathtt{i}}$ for its canonical position, and $\mu_{\mathtt{i}}^t$ for its deformed position at timestep t as predicted by DMP.

Coarse isometry. This loss prevents global geometry drift and maintains the overall object shape. We randomly sample 1% of all Gaussians from $\mathcal{G}_{\text{can}}$, forming a sparse subset $\mathcal{G}_{\text{crs}}$. For each Gaussian $g_{\mathtt{i}} \in \mathcal{G}_{\text{crs}}$, we enforce that distances to its nearest neighbors $\mathcal{N}_{\text{crs}}(g_{\mathtt{i}})$ in $\mathcal{G}_{\text{crs}}$ are preserved after deformation [41]:

$$\mathcal{L}_{\text{crs\_iso}}^t \coloneqq \sum_{\substack{g_{\mathtt{i}} \in \mathcal{G}_{\text{crs}} \\ g_{\mathtt{j}} \in \mathcal{N}_{\text{crs}}(g_{\mathtt{i}})}} \frac{w_{\mathtt{ij}}}{|\mathcal{G}_{\text{crs}}|} \left| \|\mu_{\mathtt{i}} - \mu_{\mathtt{j}}\|_2 - \|\mu_{\mathtt{i}}^t - \mu_{\mathtt{j}}^t\|_2 \right| \,, \tag{9}$$

where $w_{\mathtt{ij}} \coloneqq \text{cos-sim}(f_{\mathtt{i}}, f_{\mathtt{j}})$ weights the constraint by the DINO feature similarity between Gaussians $g_{\mathtt{i}}$ and $g_{\mathtt{j}}$, allowing semantically similar and nearby regions to deform together.

Dense isometry. This loss enforces local surface rigidity across all canonical Gaussians. For each Gaussian $g_{\mathtt{i}} \in \mathcal{G}_{\text{can}}$, we penalize distance changes to its immediate neighbors $\mathcal{N}(g_{\mathtt{i}})$:

$$\mathcal{L}_{\text{dense\_iso}}^t \coloneqq \sum_{\substack{g_{\mathtt{i}} \in \mathcal{G}_{\text{can}} \\ g_{\mathtt{j}} \in \mathcal{N}(g_{\mathtt{i}})}} \frac{w'_{\mathtt{ij}}}{|\mathcal{G}_{\text{can}}|} \left| \|\mu_{\mathtt{i}} - \mu_{\mathtt{j}}\|_2 - \|\mu_{\mathtt{i}}^t - \mu_{\mathtt{j}}^t\|_2 \right| \,, \tag{10}$$

Table 1: *Quantitative evaluation.* We evaluate our method on real-world (Panoptic Studio) and synthetic (Truebones) benchmarks against SOTA monocular dynamic reconstruction methods, as discussed in Sec. 4.4. All competitors are adapted to the pre-scan-plus-monocular setting and are conditioned on the same preprocessed inputs of 2D tracks, depth, and camera estimates. Our method achieves SOTA performance over optimization and generative-based competitors in all metrics: image reconstruction quality (PSNR, SSIM), perceptual quality (LPIPS), and semantic consistency (CLIP).

	Panoptic Studio				Truebones			
Method	**mPSNR ↑**	**mSSIM ↑**	**mLPIPS ↓**	**CLIP ↑**	**mPSNR ↑**	**mSSIM ↑**	**mLPIPS ↓**	**CLIP ↑**
HiMoR [35]	18.403	0.415	0.252	0.902	21.544	0.496	0.435	0.845
OriGS [66]	17.707	0.418	0.289	0.888	20.693	0.486	0.512	0.796
Cog-NVS [4]	16.904	0.401	0.329	0.876	21.062	0.495	0.486	0.811
Ours	**19.414**	**0.447**	**0.220**	**0.929**	**21.756**	**0.505**	**0.378**	**0.876**

where $w'_{ij} := \exp(-\beta\|\mu_i - \mu_j\|_2^2)$ assigns higher weights to closer neighbors, with β controlling the spatial falloff (see Appendix A.4 for the value).

Rigidity Loss. To further regularize the deformation field, we encourage DMP to predict similar rotations for spatially close Gaussians. For each Gaussian $g_i \in \mathcal{G}_{crs}$, we penalize rotation differences with its neighbors:

$$\mathcal{L}^t_{rigid} := \sum_{\substack{g_i \in \mathcal{G}_{crs} \\ g_j \in \mathcal{N}_{crs}(g_i)}} \frac{w_{ij}}{|\mathcal{G}_{crs}|} \|q^t_i - q^t_j\|_1 \,. \tag{11}$$

Final Objective. For each timestep t, we define the total loss:

$$\begin{aligned}\mathcal{L}^t := \mathcal{L}^t_{photo} + \lambda_{track}\mathcal{L}^t_{track} + \lambda_{depth}\mathcal{L}^t_{depth} + \lambda_{reproj}\mathcal{L}^t_{reproj} \\ + \lambda_{crs\_iso}\mathcal{L}^t_{crs\_iso} + \lambda_{dense\_iso}\mathcal{L}^t_{dense\_iso} + \lambda_{rigid}\mathcal{L}^t_{rigid} \,,\end{aligned} \tag{12}$$

where λ_* denote the relative weights between loss terms. We use a fixed set of λ_* across all experiments (see Appendix A.4 for details). Our full objective averages over all timesteps: $\mathcal{L} := \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}^t$.

4 Experiments

4.1 Implementation details

Datasets. We evaluate our method on two datasets: the Panoptic Studio dataset [21] and a synthetic dynamic animals dataset that we generate using Truebones [53]. For Panoptic Studio, we use sequences from the sports subset containing diverse single-subject human motions, including juggling, softball, tennis and boxes. Each sequence contains 150 frames at 30 FPS captured by 31 synchronized HD cameras. We simulate the monocular setting by using a single selected input view for training and evaluating on the remaining 30 held-out cameras. For the pre-scan, we use all the cameras from the first timestep. For Truebones, since the data is synthetic, we render images and ground-truth depths from known camera parameters for 7 synthetic animated characters for

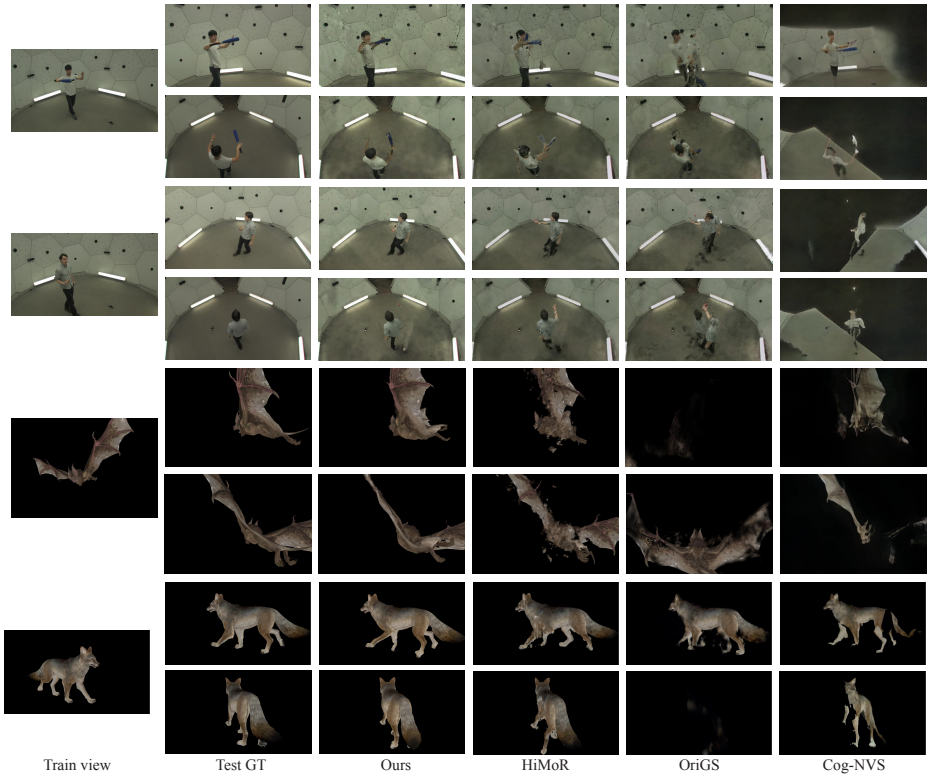

Fig. 4: *Qualitative results.* The first column depicts the training view from the monocular sequence; the second column depicts the ground-truth testing view at the same timestep. Novel views are selected at extreme angles to evaluate the completeness of dynamic 3D reconstructions. Our method drastically outperforms the baselines in maintaining a consistent object geometry and sharp appearance, while accurately modeling the scene dynamics. See our website for video results.

the pre-scan, the dynamic monocular sequence, and the testing views, allowing us to validate our approach under controlled conditions. For 3D tracking evaluation on Panoptic Studio, we use annotations from TAP-Vid-3D [26]. For Truebones, we extract the ground-truth tracks from dynamic mesh vertices. See Appendix A.6 for more evaluation and benchmarks details.

Preprocessing. We obtain FG masks using SAM2 [49]; dense 2D point tracking supervision with AllTracker [15]. For casual videos, depth and cameras are estimated using ViPE [18]. For TrueBones, we use the GT cameras and rendered GT depth. For Panoptic Studio, we render stereo cameras for the dynamic sequence using the reconstructions of [41] and obtain per-frame depth using [61].

See Appendix for architecture, hyperparameters and other implementation details.

4.2 Evaluation metrics

We evaluate and compare the performance of DRoPS in two primary tasks: (i) novel view synthesis quality, and (ii) long-range 3D tracking accuracy. For (i), we use standard image quality metrics: PSNR (Peak Signal-to-Noise Ratio) for pixel-level accuracy, SSIM [60] for structural similarity, and LPIPS [71] for perceptual quality using deep features. Additionally, we compute CLIP score [48] to measure high-level semantic consistency between rendered and ground-truth views. We measure only the FG reconstruction quality in the first 3 metrics by using ground-truth segmentation masks. For 3D tracking evaluation, we use metrics from [59], reporting the 3D end-point-error (EPE), and the fraction of points that fall within 5cm and 10cm thresholds of the ground-truth 3D positions, denoted as $\delta_{0.05}$ and $\delta_{0.1}$, respectively.

4.3 Baselines

We compare against two categories of methods: (1) Optimization-based approaches including HiMoR [35], which extends [59] with hierarchical motion representations, and OriGS [66], which utilizes orientation fields and motion scaffolds from [27], and (2) a generative method Cog-NVS [4], which uses a "warp-inpaint" approach with video diffusion models for novel view synthesis. The pre-scan is provided to all baselines within the monocular video as a subsequence preceding the dynamic motion. For a fair comparison, all methods are conditioned on the same processed data as ours, including 2D tracks, depths, and camera poses.

4.4 Results

We evaluate our method on the Panoptic Studio and Truebones datasets, comparing it against state-of-the-art optimization-based (HiMoR [35], OriGS [66]) and generative (Cog-NVS [4]) methods. Importantly, to assess the complete dynamic reconstruction of the methods, we select novel views at extreme angles that differ significantly from the input training views.

Quantitative Comparison. As shown in Tab. 1, our method significantly outperforms all competitors across all reconstruction metrics on both datasets. On the real-world Panoptic Studio dataset, which presents highly complex and non-rigid deformations, we achieve a substantial improvement over the nearest competitor (HiMoR) in both photometric reconstruction quality (over 1.0 dB in PSNR and about 13% reduction in LPIPS error) and semantic consistency in CLIP score. The performance gap is larger on the real-world data, highlighting our model's superior ability to handle the complexity and the noise of such captures. In Tab. 2, we evaluate the long-range 3D tracking performance of DRoPS compared to SOTA monocular reconstruction methods which output track estimates. As seen, DRoPS outperforms both competitors in all tracking metrics. See our website for reconstruction videos and tracking visualizations.

Qualitative Comparison. Visual comparisons in Fig. 4 further demonstrate DRoPS's superior performance over the baselines. Our method maintains superior geometric consistency and models complex motions with high fidelity,

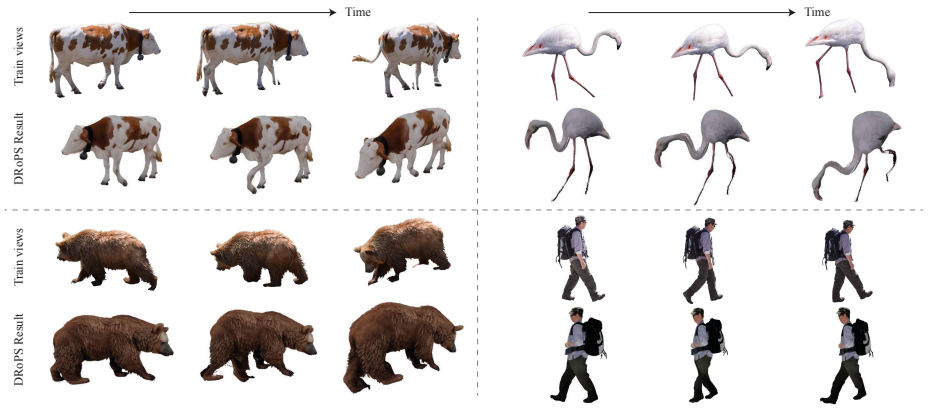

Fig. 5: DRoPS achieves high-quality dynamic 3D reconstruction on *in-the-wild monocular videos* by generating the pre-scan with an image-to-3D model [68]. Our novel views are rendered from viewpoints that differ significantly from those in the training set. See our website for full video and additional results.

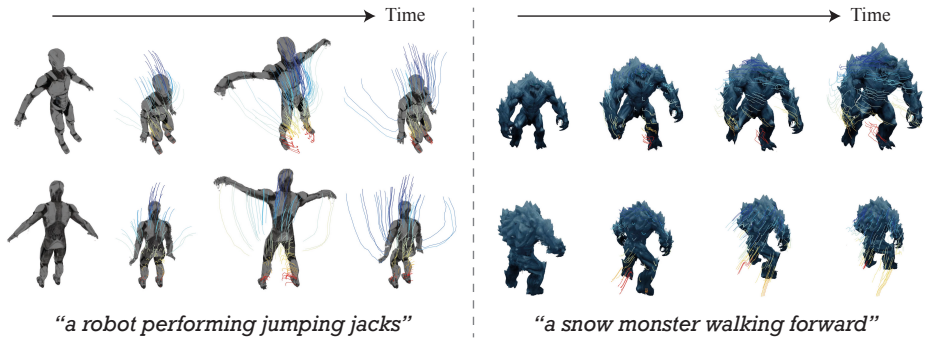

Fig. 6: *Text-to-4D application.* Each column and row correspond to a fixed timestep and viewpoint, respectively. The colored lines visualize the 3D trajectories emerging from our representation. See our website for videos.

preserving sharp textures even when rendered from extreme novel views. In contrast, competitors frequently suffer from severe artifacts, such as blurry textures, missing or extremely distorted geometry, and highly inaccurate motion (see our website for videos). Notably, even though we apply test-time fine-tuning on CogNVS for each sequence, it fails to plausibly inpaint the scene under the extreme warping required for our target novel viewpoints. Our method succeeds in leveraging the pre-scan as a geometric prior through its persistent, surface-aligned modeling of primitives and strong regularization from DMP. Crucially, despite providing the pre-scan to the competitors as a part of the monocular video, they fail to effectively exploit this explicit prior.

4.5 Additional Results

In-The-Wild Videos. Our method can operate on casual, fully monocular videos where the pre-scan is not captured by the camera. To obtain the pre-scan, we apply an image-to-3D model [68] on the first frame. As demonstrated in Fig. 5, DRoPS achieves high-quality, complete reconstruction of dynamic monocular videos, capturing the motion and appearance of the objects with high fidelity. See our website for full video visualizations and additional results.

Text-to-4D. DRoPS can be used for Text-to-4D generation: given a text prompt describing a dynamic subject, we generate a video using an off-the-shelf text-to-video model. We then use the "In-The-Wild Videos" approach for dynamic 3D reconstruction. We demonstrate our Text-to-4D results in Fig. 6.

Additional Comparison to DreamMesh4D. In Fig. 7, we provide a qualitative comparison against DM4D [34]. We evaluate DM4D both with and without a pre-scan. While DM4D suffers from distorted geometry, inaccurate motion estimation, and significant visual artifacts, DRoPS successfully leverages the pre-scan prior to achieve accurate, consistent, and realistic reconstructions.

4.6 Ablations

We ablate the key components of our design on the Panoptic Studio dataset, as demonstrated in Tab. 3 and Fig. 8. Notably, ablating each component results in a performance drop across all metrics, demonstrating the effectiveness of each component in the overall framework.

The most impactful component of DRoPS is the CNN-based motion parametrization with Deep Motion Prior. Removing DMP entirely ("w/o DMP") and directly optimizing the per-Gaussian deformation grid drastically drops in performance across all metrics. Replacing the CNN with an MLP ("w/o DMP, w/ MLP") improves over direct optimization. However, its performance remains significantly inferior to our full method. This demonstrates that the spatial inductive bias of CNNs is crucial for regularizing the ill-posed monocular reconstruction problem and for accurately capturing the scene motion.

The coarse isometry loss $\mathcal{L}_{\text{crs\_iso}}$ has a substantial impact on geometry preservation, as "w/o $\mathcal{L}_{\text{crs\_iso}}$" significantly harms all metrics. Interestingly, the dense

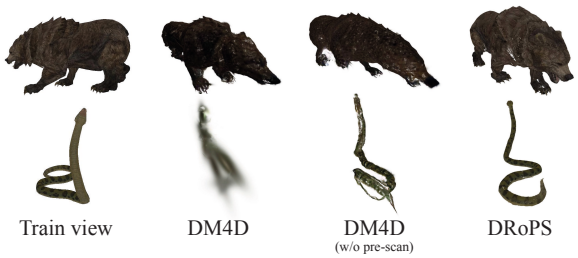

Fig. 7: We additionally compare to DM4D [34], with and without providing them the pre-scan. DM4D results in distorted geometry, inaccurate motion, and visual artifacts. In contrast, DRoPS achieves accurate, consistent and realistic reconstructions.

Table 2: Tracking Performance Comparison. We evaluate the tracking accuracy of DRoPS against state-of-the-art monocular reconstruction methods using End-to-Point Error (EPE) and the fraction of points with error below 0.05m ($\delta_{0.05}$) and 0.1m ($\delta_{0.1}$). Our method consistently outperforms baselines across both datasets.

	Panoptic Studio			Truebones		
Method	**EPE** ↓	$\boldsymbol{\delta_{0.05}}$ ↑	$\boldsymbol{\delta_{0.1}}$ ↑	**EPE** ↓	$\boldsymbol{\delta_{0.05}}$ ↑	$\boldsymbol{\delta_{0.1}}$ ↑
HiMoR [35]	0.139	0.335	0.564	0.078	0.609	0.770
OriGS [66]	0.186	0.206	0.419	0.245	0.244	0.368
Ours	**0.099**	**0.563**	**0.743**	**0.070**	**0.691**	**0.842**

isometry loss $\mathcal{L}_{\text{dense\_iso}}$ has a relatively small effect. This highlights the strength of CNN's inductive bias to regularize and implicitly correlate local motion.

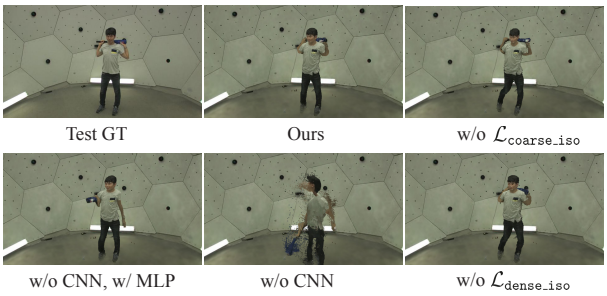

Fig. 8: *Ablation Results.* Removing DMP ("w/o CNN") overfits to the training view, failing to capture the motion and geometry. Replacing the CNN with an MLP improves over it, but still suffers from geometry distortion and inaccurate motion. $\mathcal{L}_{\text{crs\_iso}}$ is crucial for global geometry preservation, while $\mathcal{L}_{\text{dense\_iso}}$ has a relatively small impact since the inductive bias of DMP introduces strong local motion correlation.

Table 3: *Ablation Study on Panoptic Studio.* Each component's ablation results in a performance drop, highlighting their contribution. The most impactful design components are Deep Motion Prior and $\mathcal{L}_{\text{crs\_iso}}$.

Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**CLIP** ↑	**EPE** ↓	$\boldsymbol{\delta_{0.05}}$ ↑	$\boldsymbol{\delta_{0.1}}$ ↑
w/o DMP	18.289	0.409	0.313	0.9	0.274	0.131	0.267
w/o DMP, w/ MLP	18.853	0.435	0.236	0.919	0.172	0.231	0.433
w/o $\mathcal{L}_{\text{crs\_iso}}$	18.694	0.426	0.246	0.917	0.156	0.331	0.547
w/o $\mathcal{L}_{\text{dense\_iso}}$	19.218	0.433	0.227	0.926	0.157	0.346	0.558
w/o $\mathcal{L}_{\text{rigid}}$	19.257	0.440	0.223	0.927	0.147	0.383	0.599
Ours	**19.414**	**0.447**	**0.220**	**0.929**	**0.099**	**0.563**	**0.743**

5 Discussion and Conclusions

In this work, we have introduced a novel approach for complete dynamic 3D reconstruction from monocular video with a static pre-scan, enabling high-quality novel view synthesis. Our key innovations include: surface-aligned Gaussians organized into pixel grids for persistent surface representation, and a CNN-based motion parameterization that provides strong implicit regularization through its spatial inductive bias.

While DRoPS achieves excellent results, it is not without limitations. First, it currently handles single foreground subjects; extending to multi-object scenes with independent motions would require additional segmentation and per-object motion modeling. Second, DRoPS cannot handle transparencies or increasing topologies (*e.g.* fire). Third, our method relies on 2D point trackers and depth estimators for supervision, inheriting any errors from these upstream predictions. We believe these limitations are seeds for exciting future research directions.

We demonstrated the strengths of DRoPS through extensive experiments and evaluations. We showed that DRoPS significantly outperforms prior SOTA reconstruction methods in 3D tracking and novel-view synthesis, including from extreme novel views where geometric consistency is crucial. Our method offers avenues for applications beyond novel view synthesis, such as 3D animation, motion analysis, and content creation from text or casual video capture.

References

1. Attal, B., Huang, J.B., Richardt, C., Zollhoefer, M., Kopf, J., O'Toole, M., Kim, C.: HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling. In: CVPR (2023)
2. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
3. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024)
4. Chen, K., Khurana, T., Ramanan, D.: Reconstruct, inpaint, test-time finetune: Dynamic novel-view synthesis from monocular videos. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
5. Cheng, Z., Gadelha, M., Maji, S., Sheldon, D.: Bayesian deep image prior for image reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
6. Dai, P., Xu, J., Xie, W., Liu, X., Wang, H., Xu, W.: High-quality surface reconstruction using gaussian surfels. In: ACM SIGGRAPH 2024 Conference Papers. Association for Computing Machinery (2024)
7. Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: Tap-vid: A benchmark for tracking any point. In: Neural Information Processing Systems (NeurIPS) (2022)
8. Doersch, C., Yang, Y., Veber, M., Gupta, D., Markeeva, L., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: IEEE International Conference on Computer Vision (ICCV) (2023)
9. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**, 381–395 (1981), `https://api.semanticscholar.org/CorpusID:972888`
10. Fridovich-Keil, S., Meanti, G., Warburg, F., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
11. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE International Conference on Computer Vision (2021)
12. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Monocular dynamic view synthesis: A reality check. In: Neural Information Processing Systems (NeurIPS) (2022)

13. Guédon, A., Lepetit, V.: Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. CVPR (2024)
14. Harley, A.W., Fang, Z., Fragkiadaki, K.: Particle video revisited: Tracking through occlusions using point trajectories. In: European Conference on Computer Vision (ECCV) (2022)
15. Harley, A.W., You, Y., Sun, X., Zheng, Y., Raghuraman, N., Gu, Y., Liang, S., Chu, W.H., Dave, A., Tokmakov, P., You, S., Ambrus, R., Fragkiadaki, K., Guibas, L.J.: AllTracker: Efficient dense point tracking at high resolution. In: ICCV (2025)
16. Heckel, R., Hand, P.: Deep decoder: Concise image representations from untrained non-convolutional networks. In: International Conference on Learning Representations (ICLR) (2019)
17. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: SIGGRAPH 2024 Conference Papers. Association for Computing Machinery (2024). https://doi.org/10.1145/3641519.3657428
18. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixe, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. In: NVIDIA Research Whitepapers arXiv:2508.10934 (2025)
19. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. arXiv preprint arXiv:2312.14937 (2023)
20. Isik, M., Runz, M., Georgopoulos, M., Khakhulin, T., Starber, J., Agapito, L., Niesner, M.: Humanrf: High-fidelity neural radiance fields for humans in motion. In: ACM SIGGRAPH (2023)
21. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: IEEE International Conference on Computer Vision (ICCV) (2015)
22. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: arXiv preprint arXiv:2307.07635 (2023)
23. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
24. Khatib, F., Moran, D., Trostianetsky, G., Kasten, Y., Galun, M., Basri, R.: Generalizable visual localization for gaussian splatting scene representations. 2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW) pp. 178–189 (2025), https://api.semanticscholar.org/CorpusID:280711889
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), https://api.semanticscholar.org/CorpusID:6628106
26. Koppula, S., Rocco, I., Yang, Y., Heyward, J., Carreira, J., Zisserman, A., Brostow, G., Doersch, C.: Tapvid-3d: A benchmark for tracking any point in 3d. In: NeurIPS (2024)
27. Lei, J., Weng, Y., Harley, A., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. arXiv preprint arXiv:2405.17421 (2024)
28. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r (2024)
29. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: Tava: Template-free animatable volumetric actors. In: European Conference on Computer Vision (ECCV) (2022)
30. Li, T., Slavcheva, M., Zollhöfer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Lv, Z.: Neural 3d video synthesis from multi-view video (2022)

31. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8508–8520 (June 2024)
32. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
33. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: Dynibar: Neural dynamic image-based rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
34. Li, Z., Chen, Y., Liu, P.: Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
35. Liang, Y., Xu, T., Kikuchi, Y.: HiMoR: Monocular deformable gaussian reconstruction with hierarchical motion representation. In: CVPR (2025)
36. Lin, H., Peng, S., Xu, Z., Yan, Y., Shuai, Q., Bao, H., Zhou, X.: Efficient neural radiance fields for interactive free-viewpoint video. In: SIGGRAPH Asia Conference Proceedings (2022)
37. Liu, C., Chen, S., Bhalgat, Y.S., HU, S., Cheng, M., Wang, Z., Prisacariu, V.A., Braud, T.: GS-CPR: Efficient camera pose refinement via 3d gaussian splatting. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=mP7uV59iJM
38. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: The European Conference on Computer Vision (ECCV) (2018)
39. Liu, G., Shih, K.J., Wang, T.C., Reda, F.A., Sapra, K., Yu, Z., Tao, A., Catanzaro, B.: Partial convolution based padding. In: arXiv preprint arXiv:1811.11718 (2018)
40. Lombardi, S., Simon, T., Saragih, J.M., Schwartz, G., Lehrmann, A.M., Sheikh, Y.: Neural volumes. ACM Transactions on Graphics (TOG) **38**, 1 – 14 (2019), https://api.semanticscholar.org/CorpusID:195068954
41. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: International Conference on 3D Vision (3DV) (2024)
42. Mihajlovic, M., Prokudin, S., Tang, S., Maier, R., Bogo, F., Tung, T., Boyer, E.: SplatFields: Neural gaussian splats for sparse 3d and 4d reconstruction. In: European Conference on Computer Vision (ECCV). Springer (2024)
43. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV) (2020)
44. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.Y., Xu, H., Sharma, V., Li, S.W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2023)
45. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: IEEE International Conference on Computer Vision (ICCV) (2021)
46. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. In: ACM SIGGRAPH Asia (2021)

47. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
48. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML) (2021)
49. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolber, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
50. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. Springer (2015)
51. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
52. Seidenschwarz, J., Zhou, Q., Duisterhof, B.P., Ramanan, D., Leal-Taixé, L.: Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. 3DV (2025)
53. Truebones Motions Animation Studios: Truebones. `https://truebones.gumroad.com/` (2025), accessed: 2025-01-15
54. Tumanyan, N., Singer, A., Bagon, S., Dekel, T.: Dino-tracker: Taming dino for self-supervised point tracking in a single video (2024)
55. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Deep image prior. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
56. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
57. Wang, L., Hu, Q., He, Q., Wang, Z., Yu, J., Tuytelaars, T., Xu, L., Wu, M.: Neural residual radiance fields for streamably free-viewpoint videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 76–87 (June 2023)
58. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: IEEE International Conference on Computer Vision (ICCV) (2023)
59. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: International Conference on Computer Vision (ICCV) (2025)
60. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). `https://doi.org/10.1109/TIP.2003.819861`
61. Wen, B., Teed, Z., Leung, H., Dunn, G., De Nardi, R., Kautz, J.: Foundationstereo: Zero-shot stereo matching. arXiv preprint arXiv:2501.09898 (2025)
62. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: Humannerf: Free-viewpoint rendering of moving people from monocular video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
63. Williams, F., Schneider, T., Silva, C., Zorin, D., Bruna, J., Panozzo, D.: Deep geometric prior for surface reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
64. Wolf, Y., Bracha, A., Kimmel, R.: GS2Mesh: Surface reconstruction from Gaussian splatting via novel stereo views. In: European Conference on Computer Vision (ECCV) (2024)

65. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
66. Wu, J., Tao, J., Wang, H., Liu, G., Kompella, R.R., Yan, Y.: Orientation-anchored hyper-gaussian for 4d reconstruction from casual videos. In: NeurIPS (2025)
67. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
68. Xiang, J., Chen, X., Xu, S., Wang, R., Lv, Z., Deng, Y., Zhu, H., Dong, Y., Zhao, H., Yuan, N.J., Yang, J.: Native and compact structured latents for 3d generation. Tech report (2025)
69. Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H.: Banmo: Building animatable 3d neural models from many casual videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
70. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. arXiv preprint arXiv:2309.13101 (2023)
71. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
72. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024)

A Appendix

A.1 Architecture.

Our Deep Motion Prior Φ_θ is implemented as a U-Net style encoder-decoder with skip connections [50, 55], operating on the H×W pixel grids of canonical positions. The network takes the concatenation of canonical position grids $\boldsymbol{\mu}_j$, masks grid $\tilde{\mathbf{M}}_j$, and positional-encoded timestep $\gamma(t)$ (with 4 frequencies, spatially broadcasted and concatenated at each (h, w)) as input, and outputs 7-channel grids representing quaternion rotations (4 channels) and translations (3 channels). We use 4 encoder blocks with feature dimensions [16, 32, 128, 128], with skip connections of 1x1 convolutions of feature dimension 4. The output of the skip connections are concatenated with the corresponding decoder block's output channels. Each encoder block consists of a 3x3 convolutional layer, followed by a BatchNorm and a LeakyReLU activation. The decoder blocks are symmetrical to the encoder blocks, using nearest neighbor interpolation for upsampling. Crucially, Partial Convolutions [38, 39] are used to prevent empty masked regions from affecting the output.

A.2 Generated 3D Model Alignment

Image-to-3D generators (such as TRELLIS2) typically output meshes in a canonical pose, e.g. always facing in the +z direction, and are agnostic to the true image camera position relative to the 3D object. Since our method assumes the pre-scan and the first dynamic frame to be aligned, as discussed in Sec. 3.1, we align and scale TRELLIS2 mesh to match the first video frame as follows. Given the initially generated 3D model, we start from placing virtual cameras on a hemisphere surrounding the model. The renders of the mesh on each camera are processed by VGGT [56] along with the first video frame. We then sample the virtual camera with the most similar extrinsics to the first video frame estimated by VGGT. This serves as an initialization for our localization task. We then use MASt3R [28] to find 2D correspondences between the first video frame and the sampled virtual camera render. For each matched 2D point on the virtual render, we find its corresponding 3D point on the mesh by rendering the mesh depth on the camera and back-projecting. This yields 2D-to-3D correspondences from the first video frame to the 3D model points, and we use Perspective-n-Points algorithm (PnP) [9,51] to estimate the 6-DOF transformation of the mesh w.r.t. the video frame [24,37]. Lastly, we align the mesh scale to the ViPE depth scale by the median depth ratio.

A.3 Deduplication Masks

As discussed in Sec. 3.2, we apply de-duplication masks on the output of DMP to ensure each 3D region is represented by a unique primitive across virtual views $\tilde{\mathbf{T}}_j$. We start from back-projecting all Gaussians from a chosen view. Then, for each next view j, we render the depth and the transmittance of existing back-projected Gaussians to $\tilde{\mathbf{T}}_j$. We then mark the pixels in j "empty" if: (a) the rendered transmittance is below 0.1, or (b) the difference between the rendered depth and $\tilde{\mathbf{D}}_j$ is larger than a threshold 0.01. This produces a set of

de-duplication masks for each view $\tilde{\mathbf{T}}_j$ depending on the starting view. In total, we use 8 virtual cameras, and get the masks for each possible starting view, obtaining 8 sets of de-duplication masks. We uniformly sample a single mask set for each training iteration.

A.4 Hyperparameters

We train our model using the Adam optimizer [25] with a learning rate $2 \cdot 10^{-3}$ for 50,000 iterations with batch size 1. Training for a video with 150 frames takes approximately 10 hours on a single NVIDIA A100 GPU. In Eq. 2, Eq. 5 and Eq. 12, we use the following weights in all our experiments: $\lambda_{\text{SSIM}} = 0.25, \lambda_{\text{TV}} = 0.1, \lambda_{\text{track}} = 1.0, \lambda_{\text{depth}} = 100.0, \lambda_{\text{reproj}} = 100.0, \lambda_{\text{crs\_iso}} = 10.0, \lambda_{\text{dense\_iso}} = 1.0, \lambda_{\text{rigid}} = 10.0$. We set the spatial falloff parameter $\beta = 2000$ for $\mathcal{L}_{\text{dense\_iso}}$. In the coarse isometry loss, the number of nearest-neighbors in $\mathcal{N}_{\text{crs}}$ is set to $0.01 \cdot |\mathcal{G}_{\text{crs}}|$, while in the dense isometry loss, we use 200 nearest-neighbors in $\mathcal{N}$.

A.5 Truebones Benchmark

For the Truebones benchmark, we select 7 animations of highly articulated animals from their dynamic mesh dataset, which are labeled as: "Bat/AttackBite" (60 frames), "Anaconda/Strike" (140 frames), "BrownBear/Rise" (91 frames), "Bear/WalkForward" (51 frames), "Coyote/Walking" (69 frames), "Camel/Restless" (139 frames), "Camel/Run" (61 frames). We create 150 pre-scan cameras around the object in the first frame, followed by cameras for the monocular dynamic sequence, which are smoothly moving around the object while being directed towards its center at every timestep. For test views, we place 4 static cameras facing each side of the object (front, back, left, right). All videos are rendered at 1024x1024 resoltion and 30fps. See our website for visualizations of training and testing videos.

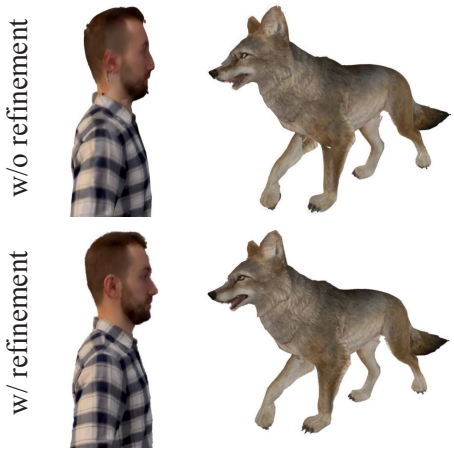

Fig. 9: *Pre-scan Refinement* optimization (Sec. 3.1) significantly improves high-frequency details in shape and texture, e.g. the geometry of the person's face, the coyote's fur and nose shape. Renders are obtained from test views.

A.6 Tracking Evaluation Details

We evaluate 3D tracking performance in both TAP-Vid-3D and our Truebones benchmark by querying each track directly at its 3D query point and tracking both forward and backward in time. Specifically, given a set of dynamic Gaussians and a 3D query point, we sample the nearest Gaussian center to the query point, and output the Gaussian center at all frames as the long-range 3D track. To assess the complete tracking and geometric ability of the reconstruction models, we measure tracking accuracy in both occluded and visible frames.

For our Truebones 3D tracking benchmark, we take the 3D positions of all vertices for each dynamic mesh sequence as the ground-truth 3D tracks. In total, we extract 20,987 tracks from 7 animations. For all methods, we query and estimate each track at its first frame location.

A.7 Canonical Depth Refinement

As discussed in Sec. 3.1, the depth estimates of FoundationStereo serve as an excellent initialization for the canonical model, but lack high-frequency details. Fig. 9 shows the canonical Gaussians before and after the refinement optimization. As seen, the refinement optimization noticeably improves the high-frequency details in the shape and texture.