# ManiSplat: Manipulation Trajectory Synthesis from Monocular Video via Decoupled 3D Gaussian Splatting

**Wenhao Hu**[1,2*], **Haonan Zhou**[1], **Liu Liu**[2†], **Yun Du**[2], **Xinjie Wang**[2], **Ziang Li**[2], **Zhizhong Su**[2] and **Gaoang Wang**[1‡]

[1]Zhejiang University
[2]Horizon Robotics

## Abstract

Reconstructing dynamic and interactive 3D scenes from real-world observations remains a fundamental challenge in computer vision and robotics. While recent advances in 3D Gaussian Splatting have enabled high-fidelity static reconstruction, extending it to interactive environments with articulated robots and manipulable objects remains difficult due to complex contact interactions and abrupt pose changes. To address these challenges, we introduce ManiSplat, a unified framework that reconstructs controllable and decoupled Gaussian digital twins directly from monocular ego-view robotic videos. Our method introduces a Graph-Structured Disentangled Representation that separates the robot, objects, and background into independently optimizable Gaussian subfields organized within a scene graph. To ensure stability, we propose a Task-Oriented Spatio-Temporal Alignment module that leverages the inherent logic of manipulation tasks—alternating between Motion and Skill phases—to construct accurate pseudo-ground-truth trajectories. Finally, a joint photometric-geometric optimization ensures the reconstructed scenes are temporally coherent, physically consistent, and simulation-ready. Extensive experiments demonstrate that our approach reconstructs interaction-driven dynamic scenes with high fidelity and controllability, effectively supporting downstream robotic tasks and policy learning. The project page is available at https://whhu7.github.io/ManiSplat/.

## 1 Introduction

Reconstructing dynamic and interactive 3D scenes from real-world observations remains a fundamental challenge in computer vision and robotics. Most existing dynamic-Gaussian frameworks treat Gaussians as functions of time, implicitly encoding motion within the representation. Methods such as Deformable 3D Gaussian [Yang *et al.*, 2024], 4D-GS [Wu *et al.*, 2024], MotionGS [Zhu *et al.*, 2024], and Spacetime Gaussian [Li *et al.*, 2024b] enable high-quality dynamic rendering but lack explicit object-level disentanglement. Later works, including Ex4DGS [Lee *et al.*, 2024], HUGS [Zhou *et al.*, 2024], EgoGaussian [Zhang *et al.*, 2025a], SplitGaussian [Li *et al.*, 2025], and BézierGS [Ma *et al.*, 2025], explicitly separate static and dynamic components to improve geometric stability and temporal coherence. However, these approaches primarily target non-interactive dynamic scenes—such as driving or human motion—where dynamics arise naturally and are only observed rather than induced. They remain limited in interaction-driven scenarios, where articulated agents (e.g., robotic arms) actively manipulate objects, producing complex contact interactions and pose changes. Although IGFuse [Hu *et al.*, 2026] explores interactive Gaussian reconstruction by fusing multi-state scene scans, it relies on discrete observations with object rearrangements rather than continuous robotic manipulation videos.

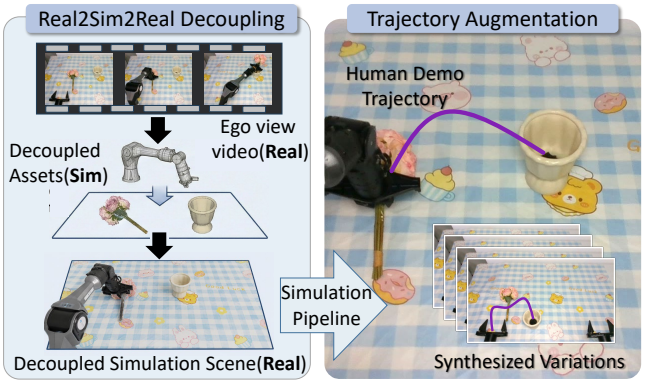


Figure 1: **High-Fidelity Real2Sim Alignment and Trajectory Augmentation. (Left)** Our framework decomposes monocular ego-view video into aligned assets (URDF robot, objects, and background). **(Right)** This enables a trajectory augmentation pipeline that synthesizes diverse variations from single demonstrations to scale up data for policy learning.

Meanwhile, the Real2Sim2Real (R2S2R) paradigm has emerged as an effective strategy to bridge real-world perception and simulation, enabling consistent transfer of sensing and control for robotic learning[Chebotar *et al.*, 2019]. Re-

[*]This work was done during an internship at Horizon Robotics.
[†]Project lead.
[‡]Corresponding author.

cent studies have incorporated 3DGS into robotic pipelines to enhance photorealism and physical consistency. For instance, Robo-GS [Lou *et al.*, 2025], RL-GSBridge [Wu *et al.*, 2025], SplatSim [Qureshi *et al.*, 2025], and RoboGSim [Li *et al.*, 2024a] couple Gaussian Splatting with physics-aware constraints to achieve differentiable rendering and physically consistent control; DREMA [Barcellona *et al.*, 2024] and RE3SIM [Han *et al.*, 2025] further construct learnable digital twins for embodied policy learning. However, most of these frameworks rely on static scene scans and pre-aligned assets, making it difficult to model robots directly from dynamically captured videos.

To address these limitations, we propose a unified framework for reconstructing interaction-driven dynamic Gaussian scenes directly from monocular ego-view robotic videos. First, we design a Graph-Structured Disentangled Representation that explicitly separates the robot arm, manipulable objects, and static background into independently optimizable Gaussian subfields within a scene graph. Second, we present a Task-Oriented Spatio-Temporal Alignment module. This module leverages the inherent logic of manipulation tasks—alternating between Motion and Skill phases—to construct accurate pseudo-ground-truth pose trajectories through Hybrid Pose Estimation and appearance alignment, bridging the domain gap while ensuring stability.

Finally, leveraging this decoupled and controllable representation, we develop a Topology-Preserving Data Augmentation pipeline. By adopting a "skill-preserving, motion-adaptive" strategy, we can apply rigid transformations to identified skill segments while autonomously re-planning motion segments. This mechanism enables the generation of large-scale, physically feasible, and spatially diverse synthetic trajectories from a single real-world demonstration, providing high-fidelity data support for downstream embodied policy learning.

In summary, our main contributions are as follows:

- We propose a unified framework for reconstructing interactive and simulation-ready Gaussian scenes from real-world monocular ego-view robotic videos.
- We design a graph-based disentangled representation system that achieves independent modeling and object-level controlled reconstruction of the robot, objects, and background.
- We introduce a task-oriented spatio-temporal alignment mechanism that significantly improves trajectory stability and rendering fidelity during interactions through hybrid pose estimation and appearance constraints.
- We develop a topology-preserving data augmentation method capable of synthesizing physically consistent augmented trajectories from a single demonstration, effectively addressing data sparsity and enhancing policy generalization.

# 2 Related Works

## 2.1 Real2Sim2Real

The Real2Sim2Real (R2S2R) paradigm bridges real-world data and simulation for robotic learning, enabling consistent transfer of perception and control. Recent efforts integrate 3D Gaussian Splatting (3DGS) into robotic pipelines to enhance photorealism and physical realism. Robo-GS [Lou *et al.*, 2025] fuses mesh geometry, Gaussian kernels, and physics attributes via a Gaussian–Mesh–Pixel binding mechanism for differentiable, physically grounded rendering. RL-GSBridge [Wu *et al.*, 2025] and SplatSim [Qureshi *et al.*, 2025] embed 3DGS within reinforcement learning frameworks, ensuring visually consistent control and reducing Sim2Real gaps. RoboGSim [Li *et al.*, 2024a], DREMA [Barcellona *et al.*, 2024], and RE3SIM [Han *et al.*, 2025] further integrate Gaussian splatting with physics simulation to construct learnable, photorealistic digital twins for policy learning. RoboSplat [Yang *et al.*, 2025] directly edits reconstructed Gaussian scenes to generate diverse robotic demonstrations. Unlike these methods that depend on static scene scans for geometry and pose alignment, our approach performs video-based dynamic disentanglement and fusion, leveraging temporal motion cues from real videos to unify appearance and pose across the real and simulated domains.

## 2.2 Dynamic Gaussian Reconstruction

Recent advances in dynamic scene modeling extend static 3D Gaussian splatting to the temporal domain. Deformable 3D Gaussian [Yang *et al.*, 2024] reconstructs dynamic scenes in a canonical space via a learned deformation field, enabling real-time, temporally smooth rendering. 4D-GS [Wu *et al.*, 2024] jointly encodes space and time through 3D Gaussians coupled with 4D neural voxels for compact, high-resolution rendering. MotionGS [Zhu *et al.*, 2024] introduces optical-flow-based motion priors to decouple camera and object motion, improving monocular reconstruction accuracy. Spacetime Gaussian [Li *et al.*, 2024b] augments 3D Gaussians with temporal opacity and motion attributes to achieve high-fidelity, view- and time-dependent rendering. MEGA [Zhang *et al.*, 2025b] and FreeTimeGS [Wang *et al.*, 2025] further improve efficiency and flexibility through compact color encoding and learnable motion functions, respectively. However, these approaches generally treat Gaussians as temporal functions and fail to disentangle individual moving objects from the scene. To address this, recent works such as Ex4DGS [Lee *et al.*, 2024], HUGS [Zhou *et al.*, 2024], EgoGaussian [Zhang *et al.*, 2025a], SplitGaussian [Li *et al.*, 2025], and BézierGS [Ma *et al.*, 2025] explicitly separate static and dynamic components, enabling more stable and consistent motion reconstruction. Nevertheless, existing methods are primarily designed for passive dynamic scenes (e.g., driving or human motion) and cannot effectively model robotic manipulation scenarios, where object trajectories are actively induced by articulated agents. In contrast, our framework explicitly models interaction-driven dynamics within Gaussian fields, bridging the gap between dynamic reconstruction and physically grounded manipulation understanding.

# 3 Preliminary

## 3.1 Gaussian Splatting

Following [Kerbl *et al.*, 2023], a 3D scene is represented as a set of Gaussian primitives $\mathcal{G} = \{\mathbf{x}, \mathbf{\Sigma}, \alpha, \mathbf{c}\}$, where

$\mathbf{x}$ denotes the 3D center position, $\mathbf{\Sigma}$ is the spatial covariance matrix, $\alpha$ is the opacity coefficient, and $\mathbf{c}$ is the RGB color vector. During rendering, each Gaussian is projected onto the 2D image plane through a differentiable $\alpha$-blending process, and the final pixel color $\mathbf{C}$ is computed by accumulating Gaussian contributions along the viewing ray as $\mathbf{C} = \sum_{i\in\mathcal{N}} \mathbf{c}_i \alpha'_i \prod_{j<i}(1-\alpha'_j)$, where $\mathcal{N}$ denotes the ordered set of Gaussians intersected by the ray.

We further define the application of a rigid transformation $T = (R, t) \in \mathrm{SE}(3)$ to a Gaussian as $T \otimes \mathcal{G} = \{R\mathbf{x} + t,\ R\mathbf{\Sigma}R^\top,\ \alpha,\ \mathbf{c}\}$, where $R$ and $t$ denote the rotation and translation components, respectively. This operation transforms the Gaussian's center and covariance into world coordinates while keeping its opacity and color unchanged.

# 4 Method

As illustrated in Fig. 2, our framework reconstructs controllable Gaussian scenes from real robotic videos and generates diverse, physically feasible demonstration data. The system consists of three core components: **Graph-Structured Disentangled Representation** (§4.1) models the robot, objects, and background as independent Gaussian subfields to achieve a structured scene representation. **Task-Oriented Spatio-Temporal Alignment** (§4.2) leverages interaction logic (Motion vs. Skill) to construct pseudo-ground-truth object poses, while performing appearance alignment to ensure the rendered Gaussians are consistent with real-world observations in both pose and visual features. **Topology-Preserving Data Augmentation** (§4.4) synthesizes new trajectory data by rigidly transforming skills and re-planning motions, significantly scaling up the demonstration data. Finally, we present our unified optimization objective in §4.3.

## 4.1 Graph-Structured Disentangled Representation

**Gaussian Scene Graph.** To strictly disentangle the scene for downstream control, we represent the world as a dynamic scene graph comprising three fundamental types of nodes: (1) The Background Node ($G_{\text{bg}}$), which contains static Gaussians representing the environment; (2) The Robot Node ($G_{\text{robot}}$), modeled as articulated Gaussian clusters driven by the URDF kinematic chain; (3) The Object Nodes ($G^v_{\text{obj}}$), representing manipulable tools or rigid entities.

At any time step $t$, the complete scene is composed via the union of these transformed subfields:

$$G(t) = G_{\text{bg}} \cup G_{\text{robot}}(t) \cup \{G^v_{\text{obj}}(t)\}_v \tag{1}$$

**Robot Node.** The robot geometry is explicitly driven by its kinematic state. Given joint angles $\mathbf{q}_t \in \mathbb{R}^n$, the Gaussian primitives attached to link $l$, denoted as $\bar{G}^{(l)}_{\text{robot}}$, are transformed to world space via forward kinematics $T_{\text{base}\to l}(\mathbf{q}_t)$:

$$G^{(l)}_{\text{robot}}(t) = T_{\text{base}\to l}(\mathbf{q}_t) \otimes \bar{G}^{(l)}_{\text{robot}}, \tag{2}$$

where $\otimes$ denotes the application of an $\mathrm{SE}(3)$ transformation to the Gaussian centers and covariances.

**Object Nodes.** Each object node $v$ maintains a canonical Gaussian set $\bar{G}^v_{\text{obj}}$ and a time-varying pose $T_v(t) \in \mathrm{SE}(3)$. The world-space representation is computed as:

$$G^v_{\text{obj}}(t) = T_v(t) \otimes \bar{G}^v_{\text{obj}}. \tag{3}$$

By decoupling object-specific Gaussians from the background, our framework enables high-fidelity rendering of object interactions and supports flexible trajectory manipulation.

## 4.2 Task-Oriented Spatio-Temporal Alignment

Accurate estimation of object pose $T_v(t)$ is critical for reconstruction. Instead of relying on noisy optical flow or heuristic smoothing, we propose a task-oriented approach that utilizes the inherent structure of manipulation tasks—alternating between free-space motion and contact-rich interaction.

**Interaction-Aware Phase Segmentation.** Following the semantic decomposition in DemoGen [Xue *et al.*, 2025], we partition the demonstration into two distinct phases based on the proximity between the robot end-effector (EE) and the target object. Let $d(t)$ denote the distance between the EE and the object center. We define a proximity threshold $\tau$ to classify the trajectory: the **Skill Phase ($S$)** occurs when $d(t) < \tau$, representing critical interactions like grasping or placing; otherwise, the segment is classified as the **Motion Phase ($M$)**, representing free-space transport or approaching. For a typical pick-and-place task, the sequence is naturally segmented as: $M_{\text{approach}} \to S_{\text{pick}} \to M_{\text{transfer}} \to S_{\text{place}} \to M_{\text{return}}$.

**Hybrid Pose Estimation.** Based on these phases, we construct a reliable pseudo-ground-truth pose trajectory, denoted as $T^{\text{pseudo}}_v(t)$, by applying phase-specific geometric constraints:

1. **Static Constraint ($M_{\text{approach}}, M_{\text{return}}$):** When the object is not involved in manipulation, it remains static in the world frame. We fix its pose to the initial or final stable state::
$$T^{\text{pseudo}}_v(t) = \begin{cases} T_v(t_{\text{init}}), & \forall t \in M_{\text{approach}}, \\ T_v(t_{\text{final}}), & \forall t \in M_{\text{return}}. \end{cases} \tag{4}$$
where $T_v(t_{\text{init}})$ denotes the initial stable pose of the object before the task starts, and $T_v(t_{\text{final}})$ denotes the final stable pose of the object after the task is completed.
2. **Attached Constraint ($M_{\text{transfer}}$):** During transport, the object is rigidly attached to the gripper. We assume the relative transformation $T_{\text{rel}}$ (captured at the end of the grasp skill) remains constant. The object pose is derived purely from robot kinematics:
$$T^{\text{pseudo}}_v(t) = T_{\text{ee}}(t) \cdot T_{\text{rel}}, \quad \forall t \in M_{\text{transfer}} \tag{5}$$
where $T_{\text{ee}}(t)$ is the accurate end-effector pose from the robot controller.
3. **Tracking-Based Constraint ($S_{\text{pick}}, S_{\text{place}}$):** During interaction, the object may undergo complex movements that are neither static nor fully rigid with the EE. To capture this, we employ a point tracker to track a set of 3D keypoints $\mathcal{P} = \{\mathbf{p}_k\}$ on the object. We solve for the optimal pose $T(t)$ that minimizes the registration error:
$$T^{\text{pseudo}}_v(t) = \underset{T\in\mathrm{SE}(3)}{\arg\min} \sum_k \|T \cdot \mathbf{p}_k(0) - \mathbf{p}_k(t)\|^2 \tag{6}$$

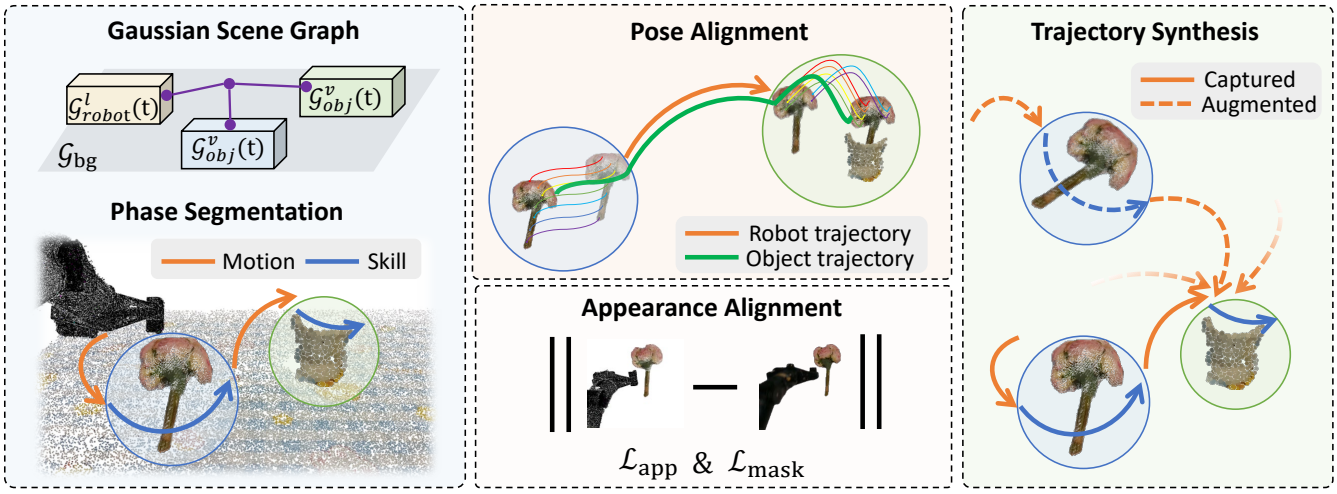


Figure 2: **Overview of the Proposed Framework. (Left)** We construct a *Gaussian Scene Graph* ($\mathcal{G}$) to disentangle the scene into independent semantic nodes: the robot $\mathcal{G}_{robot}$, the object $\mathcal{G}_{obj}$, and the static background $\mathcal{G}_{bg}$. Simultaneously, the manipulation task is segmented into a robot-centric *Motion* phase (orange, e.g., transport) and an object-centric *Skill* phase (blue, e.g., insertion). **(Middle)** To build a high-fidelity digital twin, we perform joint optimization via *Pose Alignment* and *Appearance Alignment*. **(Right)** Leveraging the decoupled structure, we perform *Trajectory Synthesis* by generating diverse augmented approach paths (dashed orange lines) that seamlessly converge into the preserved skill execution, significantly scaling up the demonstration data.

**Appearance Alignment.** Lighting and texture differences between the URDF-rendered robot arm and its real counterpart often lead to a significant appearance gap. To bridge this domain gap, we utilize SAM2 [Ravi *et al.*, 2024] to generate dynamic masks $M_{\text{dyn}}$ for the moving components. During training, appearance consistency is enforced via masked $\mathcal{L}_1$ and $\mathcal{L}_{\text{SSIM}}$ losses:

$$\mathcal{L}_{\text{app}} = \lambda_1 \mathcal{L}_1(I \cdot M_{\text{dyn}}, \hat{I} \cdot M_{\text{dyn}}) + \lambda_{\text{ssim}} \mathcal{L}_{\text{SSIM}}(I \cdot M_{\text{dyn}}, \hat{I} \cdot M_{\text{dyn}}). \tag{7}$$

To further disentangle dynamic objects from the static background, we impose a mask-consistency constraint between the rendered alpha map of dynamic Gaussians $\mathcal{O}^G_{\text{dyn}}$ and the SAM2 mask $M_{\text{dyn}}$:

$$\mathcal{L}_{\text{mask}} = \|\mathcal{O}^G_{\text{dyn}} - M_{\text{dyn}}\|_1. \tag{8}$$

Enforcing these alignment losses not only bridges the visual gap but also promotes coarse pose alignment, as visual consistency implicitly guides the geometric optimization of the scene graph.

## 4.3 Optimization

We jointly optimize the Gaussian parameters and the scene graph structure. In addition to the standard photometric loss used in vanilla 3D Gaussian Splatting [Kerbl *et al.*, 2023], which ensures pixel-level color and structural consistency, we incorporate strong geometric and appearance constraints to handle the challenges of robotic manipulation scenes.

Specifically, we introduce a pose loss $\mathcal{L}_{\text{pose}}$ to supervise the object trajectories based on our hybrid estimation:

$$\mathcal{L}_{\text{pose}} = \sum_t \|T_v(t) \ominus T_v^{\text{pseudo}}(t)\|_1, \tag{9}$$

where $\ominus$ denotes the distance metric in $\text{SE}(3)$, covering both rotation and translation.

The complete optimization objective is formulated as a weighted sum of the rendering, appearance, mask, and pose terms:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{render}} \mathcal{L}_{\text{render}} + \lambda_{\text{app}} \mathcal{L}_{\text{app}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{pose}} \mathcal{L}_{\text{pose}}. \tag{10}$$

By integrating these losses, our framework ensures that the reconstructed scene is not only visually high-fidelity but also physically consistent with the robot's kinematic constraints, facilitating effective Real2Sim2Real transfer.

## 4.4 Topology-Preserving Data Augmentation

To scale up demonstration data for policy learning, we generate synthetic trajectories by modifying the scene layout while preserving the underlying task logic. We adopt a "Skill-Preserving, Motion-Adaptive" strategy that augments object poses and ensures kinematic feasibility through autonomous re-planning.

**Skill Augmentation via Rigid Transformation.** Manipulation skills rely on the local relative pose between the gripper and the object. For an identified Skill segment $S_i$, we apply a random rigid transformation $T_{\text{aug}} \in \text{SE}(3)$ to the object's pose. To maintain the integrity of the interaction mechanics, the end-effector trajectory is subjected to the same transformation:

$$T'_{\text{ee}}(t) = T_{\text{aug}} \cdot T_{\text{ee}}(t), \quad \forall t \in S_i \tag{11}$$

This operation ensures that delicate actions, such as grasping or insertion, remain geometrically invariant and are simply transposed to a new task coordinates defined by $T_{\text{aug}}$.

**Motion Generation via Re-planning.** Motion segments serve as transitions between consecutive skills. Directly transforming the original motion trajectory often results in collisions or kinematic singularities within the augmented

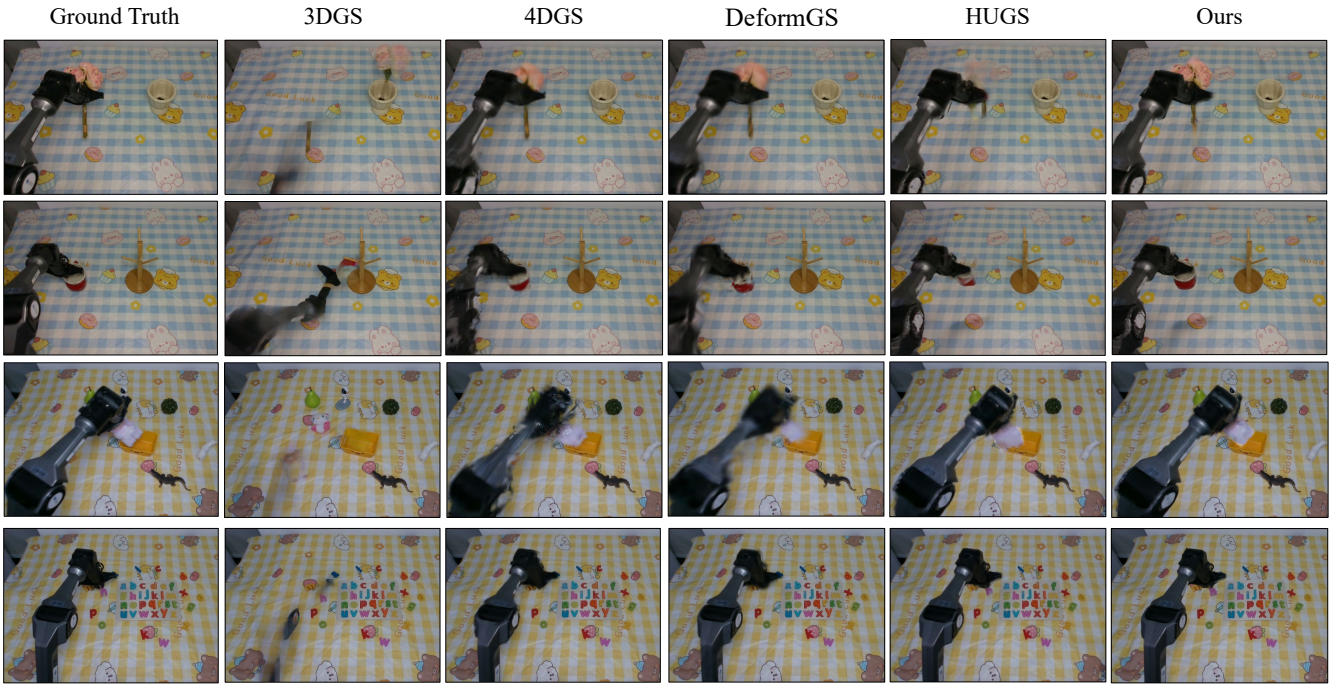


Figure 3: Qualitative comparison of dynamic reconstruction. 4DGS and DeformableGS exhibit artifacts and blurring during rapid manipulator movements due to the lack of decoupling. HUGS produces blurry object details by over-smoothing the abrupt pose changes during grasping. In contrast, our method maintains sharp boundaries and stable appearance by explicitly modeling the interaction stages.

layout. To address this, we treat each Motion segment $M_i$ as a navigation problem between the augmented skills. Let the terminal state of the preceding augmented skill be $\mathbf{q}'_{\text{start}}$ and the initial state of the subsequent augmented skill be $\mathbf{q}'_{\text{goal}}$. We invoke a motion planning framework to generate a new, collision-free trajectory that connects these states. This decoupled approach enables the robot to adaptively navigate the modified environment while strictly adhering to the core manipulation logic.

# 5 Experiments

## 5.1 Dataset

To comprehensively evaluate our method, we utilize both synthetic and real-world datasets. The synthetic data is generated using the RoboTwin simulation platform [Chen *et al.*, 2025], which provides high-fidelity rendering and precise ground truth parameters. The real-world data is captured using an AgileX dual-arm manipulator, recording complex interaction sequences in physical environments.

For the reconstruction task, we employ a frame-subsampling strategy to partition the data into training and testing sets, evaluating the novel view synthesis quality. For the pose estimation task, we rely on the synthetic ground truth pose trajectories to quantitatively assess the accuracy of our estimations.

## 5.2 Experimental Setup

**Implementation details.** The training process is conducted for a total of 30,000 iterations using the Adam optimizer. To ensure robust scene initialization, we obtain the initial object poses via GenPose++ [Zhang *et al.*, 2024]. For the segments requiring dynamic tracking, CoTracker [Karaev *et al.*, 2025] is employed to provide point-based estimation. The distance threshold $\tau$, which distinguishes between the motion and skill phases, is set to 0.2. The loss balancing hyperparameters are assigned as $\lambda_{\text{app}} = 1.0$, $\lambda_{\text{mask}} = 1.0$, and $\lambda_{\text{pose}} = 1.0$. All experiments and runtime evaluations are performed on a single NVIDIA RTX 4090 GPU.

**Baselines.** We compare our proposed method against several representative Gaussian Splatting-based frameworks, categorized by their modeling capabilities. First, we include **3DGS** [Kerbl *et al.*, 2023] as a baseline for static scene reconstruction, serving as a reference that lacks dynamic modeling capabilities. To evaluate dynamic scene modeling, we compare against **4DGS** [Wu *et al.*, 2024] and **DeformableGS** [Yang *et al.*, 2024]. While these methods effectively handle non-rigid deformations, they do not explicitly decouple the object from the environment. Finally, we examine decoupled dynamic reconstruction methods, including **HUGS** [Zhou *et al.*, 2024], which employs function fitting for pose curves. To ensure a fair comparison with the latter, we extend the dynamics function within HUGS to support full 3D spatial modeling.

## 5.3 Dynamic Reconstruction

**Qualitative Analysis.** The qualitative performance of dynamic reconstruction is illustrated in Fig. 3, while quantitative results are detailed in Tab. 1. Due to the lack of explicit

| Methods | *Dynamic* | *Decouple* | Full Image PSNR↑ | Full Image SSIM↑ | Robot PSNR↑ | Robot SSIM↑ | Object PSNR↑ | Object SSIM↑ |
|---|---|---|---|---|---|---|---|---|
| 3DGS [Kerbl *et al.*, 2023] | | | 19.93 | 0.898 | 5.36 | 0.415 | 12.57 | 0.633 |
| 4DGS[Wu *et al.*, 2024] | ✓ | | 30.68 | 0.937 | 21.37 | 0.824 | 23.41 | 0.865 |
| DeformGS[Yang *et al.*, 2024] | ✓ | | 30.24 | 0.940 | 20.20 | 0.818 | 22.96 | 0.863 |
| HUGS[Zhou *et al.*, 2024] | ✓ | ✓ | 30.70 | 0.893 | 26.07 | 0.838 | 18.90 | 0.767 |
| **Ours** | ✓ | ✓ | **32.31** | **0.951** | **26.07** | **0.838** | **26.85** | **0.872** |

Table 1: Quantitative comparison of scene reconstruction. We compare our method with state-of-the-art methods on full image, robot, and object regions.

| Method | $E_t$ (cm) ↓ | $E_R$ (°) ↓ |
|---|---|---|
| GenPose++ [Zhang *et al.*, 2024] | 0.9965 | 7.109 |
| HUGS [Zhou *et al.*, 2024] | 0.9936 | 7.004 |
| **Ours** | **0.5864** | **5.185** |

Table 2: **Quantitative comparison on RoboTwin.** We report the average translation error ($E_t$) in cm and rotation error ($E_R$) in degrees. The best results are highlighted in **bold**.

object-environment decoupling, 4DGS and DeformableGS struggle to handle the rapid motion of the robotic arm, often failing to deform correctly and resulting in significant artifacts and motion blur. While decoupled methods address this to some extent, they face distinct challenges in pose estimation. HUGS employs dynamics curve fitting to constrain motion; however, during high-degree-of-freedom interactions—such as the moment of grasping—the strict smoothing function tends to obliterate the necessary abrupt changes in object pose, similarly leading to blurred reconstruction. Our method addresses these limitations through a multi-stage pose handling strategy. In the motion phase, we leverage robot kinematics to maintain relative static attributes, while in the abrupt skill phase, we incorporate tracking supervision. This allows us to maintain stability in both pose and appearance throughout the entire interaction sequence.

**Quantitative Analysis.** Quantitatively, 3DGS yields the lowest PSNR as it lacks dynamic modeling capabilities. While 4DGS and DeformableGS improve upon this with dynamic reconstruction, their inability to decouple the scene results in poor performance when evaluating the manipulator and object separately. For fair comparison, we supply HUGS with ground-truth robot joint angles, resulting in manipulator reconstruction quality comparable to ours. However, their object reconstruction metrics are significantly lower, as HUGS suffers from over-smoothing in the presence of complex contact. Our method achieves superior performance across all metrics by accurately capturing the complex interaction dynamics.

## 5.4 Pose Estimation

**Pose Estimation Accuracy.** We evaluate pose accuracy using RoboTwin ground-truth. As shown in Tab. 2, our method achieves the lowest error, outperforming GenPose++ and HUGS. Specifically, we reduce $E_t$ to 0.5864 cm and $E_R$ to 5.185°, a significant improvement over GenPose++ (0.9965 cm, 7.109°). By leveraging kinematic data for refinement, our approach effectively mitigates estimation drift and ensures high-fidelity reconstruction.

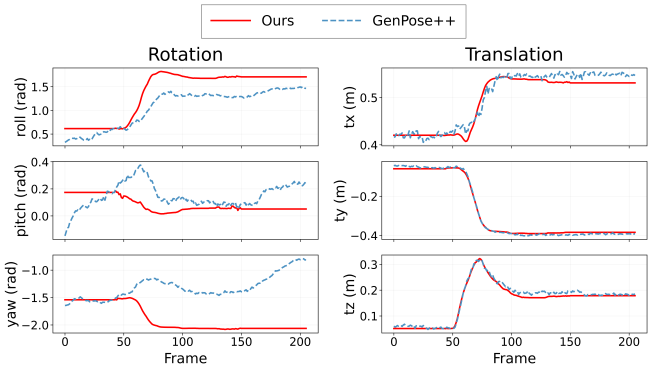


Figure 4: Qualitative Comparison of Pose Trajectories. The plots depict object rotation (left, in radians) and translation (right, in meters) across the sequence (Frame ID). Red solid lines represent our method, while blue dashed lines represent the GenPose++ baseline. Our framework achieves superior temporal stability while precisely capturing abrupt pose transitions during interaction events.

**Trajectory Analysis.** Fig. 4 visualizes real-world trajectories. Compared to the GenPose++ baseline, our method demonstrates superior temporal stability by suppressing high-frequency noise. Crucially, it accurately captures the abrupt pose transitions during gripper engagement (e.g., near Frame 50), which are typically obscured by noise in single-frame estimation methods. This ensures a stable trajectory during the motion phase while precisely reflecting discrete physical events during the skill phase.

## 5.5 Data Augmentation

Beyond reconstruction, our decoupled representation enables high-fidelity data augmentation. Figure 5 demonstrates this capability using a scene containing a flower and a vase. The first row displays the original grasping demonstration. The bottom two rows show augmented sequences where both the position and orientation of the flower are altered, while the vase remains stationary. Our method successfully generates photorealistic novel scenarios from a single source video,

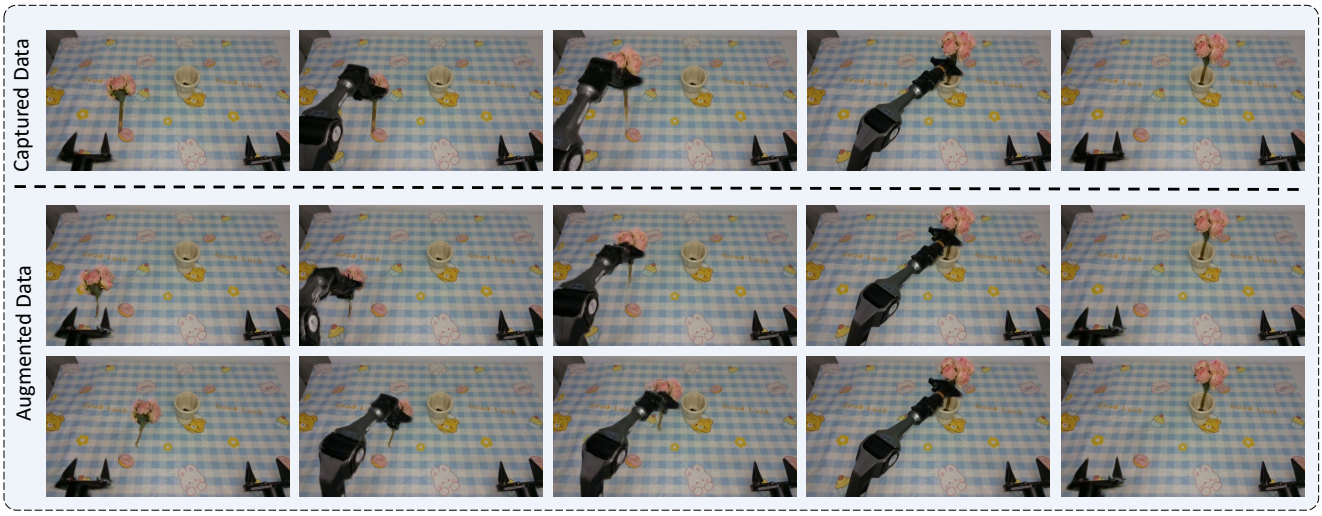


Figure 5: Data augmentation results. The first row shows the original grasping demonstration. The subsequent rows display high-fidelity augmentations where the flower and vase are spatially manipulated, demonstrating our method's ability to generate diverse training data from a single video.

| $\mathcal{L}_{\text{pose}}$ | $\mathcal{L}_{\text{app}}$ | $\mathcal{L}_{\text{mask}}$ | PSNR ↑ | SSIM ↑ |
|---|---|---|---|---|
| - | - | - | 30.08 | 0.932 |
| ✓ | - | - | 31.15 | 0.943 |
| ✓ | ✓ | - | 32.24 | 0.949 |
| ✓ | ✓ | ✓ | **32.31** | **0.951** |

Table 3: Ablation study of $\mathcal{L}_{\text{pose}}$ (Pose constraint), $\mathcal{L}_{\text{app}}$ (Appearance alignment), and $\mathcal{L}_{\text{mask}}$ (Opacity mask supervision) on the RoboTwin dataset.

preserving interaction fidelity across different spatial configurations. This capability is particularly valuable for scaling datasets for downstream tasks, such as Vision-Language-Action (VLA) model training, where diverse object poses are critical for generalization.

We conduct an ablation study to evaluate the individual contributions of the pose constraint $\mathcal{L}_{\text{pose}}$, appearance alignment loss $\mathcal{L}_{\text{app}}$, and opacity mask supervision $\mathcal{L}_{\text{mask}}$. As shown in Tab. 3, $\mathcal{L}_{\text{pose}}$ provides the most significant performance gain, as it ensures geometric stability and reduces the high-frequency jitter inherent in sparse-view reconstruction. The introduction of $\mathcal{L}_{\text{app}}$ further refines the rendering quality by enforcing photometric consistency across the sequence.

Notably, while the opacity-controlling $\mathcal{L}_{\text{mask}}$ yields a relatively marginal improvement in global metrics, it is indispensable for obtaining high-quality decoupled representations. By explicitly supervising the Gaussian opacity, $\mathcal{L}_{\text{mask}}$ effectively eliminates floaters and ensures sharp, physically plausible boundaries between the robot manipulator and the manipulated objects. This precise disentanglement is critical for downstream tasks like data augmentation, where clean object-environment separation is required for realistic recomposition.

## 6 Limitation and Future Work

Despite its effectiveness in structured scene reconstruction, our framework still has limitations. In particular, rendering quality may degrade when data augmentation introduces large spatial offsets from the original demonstration. This is mainly due to the sparse-view nature of 3D Gaussian Splatting, where Gaussians optimized from limited viewpoints may lack sufficient geometric consistency for large-viewpoint generalization.

Several directions remain for future work. First, replacing 3DGS with more geometrically constrained representations, such as 2D Gaussian Splatting (2DGS), could improve surface consistency and view synthesis quality. Second, augmentation quality could be further enhanced by incorporating multiple reference demonstrations. By fusing multi-state information from diverse interaction sequences of the same object, the framework could build a more robust scene representation and support higher-fidelity synthesis under large spatial perturbations.

## 7 Conclusion

We present a unified framework for reconstructing interactive, simulation-ready Gaussian digital twins from real-world monocular egocentric robotic videos. Our framework adopts a graph-based disentangled representation to separately model the robot, objects, and background, enabling object-level controllable reconstruction. A task-oriented spatio-temporal alignment mechanism improves trajectory stability and rendering fidelity by separating free-space motion from contact-rich interaction phases. To alleviate data sparsity, we further introduce a topology-preserving augmentation strategy that synthesizes physically consistent trajectories from a single demonstration. Overall, our approach provides a scalable pipeline for converting raw monocular robotic videos into high-fidelity digital assets.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62576308), Zhejiang Provincial Natural Science Foundation of China (No. LZ24F030005), and Fundamental Research Funds for the Central Universities (No. 226-2025-00167).

## References

[Barcellona *et al.*, 2024] Leonardo Barcellona, Andrii Zadaianchuk, Davide Allegro, Samuele Papa, Stefano Ghidoni, and Efstratios Gavves. Dream to manipulate: Compositional world models empowering robot imitation learning with imagination. *arXiv preprint arXiv:2412.14957*, 2024.

[Chebotar *et al.*, 2019] Yevgen Chebotar, Ankur Handa, Viktor Makoviychuk, Miles Macklin, Jan Issac, Nathan Ratliff, and Dieter Fox. Closing the sim-to-real loop: Adapting simulation randomization with real world experience. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8973–8979, 2019.

[Chen *et al.*, 2025] Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

[Han *et al.*, 2025] Xiaoshen Han, Minghuan Liu, Yilun Chen, Junqiu Yu, Xiaoyang Lyu, Yang Tian, Bolun Wang, Weinan Zhang, and Jiangmiao Pang. Re3sim: Generating high-fidelity simulation data via 3d-photorealistic real-to-sim for robotic manipulation. *arXiv preprint arXiv:2502.08645*, 2025.

[Hu *et al.*, 2026] Wenhao Hu, Zesheng Li, Haonan Zhou, Liu Liu, Xuexiang Wen, Zhizhong Su, Xi Li, and Gaoang Wang. Igfuse: Interactive 3d gaussian scene reconstruction via multi-scans fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4932–4940, 2026.

[Karaev *et al.*, 2025] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025.

[Kerbl *et al.*, 2023] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[Lee *et al.*, 2024] Junoh Lee, ChangYeon Won, Hyunjun Jung, Inhwan Bae, and Hae-Gon Jeon. Fully explicit dynamic gaussian splatting. *Advances in Neural Information Processing Systems*, 37:5384–5409, 2024.

[Li *et al.*, 2024a] Xinhai Li, Jialin Li, Ziheng Zhang, Rui Zhang, Fan Jia, Tiancai Wang, Haoqiang Fan, Kuo-Kun Tseng, and Ruiping Wang. Robogsim: A real2sim2real robotic gaussian splatting simulator. *arXiv preprint arXiv:2411.11839*, 2024.

[Li *et al.*, 2024b] 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.

[Li *et al.*, 2025] Jiahui Li, Shengeng Tang, Jingxuan He, Gang Huang, Zhangye Wang, Yantao Pan, and Lechao Cheng. Splitgaussian: Reconstructing dynamic scenes via visual geometry decomposition. *arXiv preprint arXiv:2508.04224*, 2025.

[Lou *et al.*, 2025] Haozhe Lou, Yurong Liu, Yike Pan, Yiran Geng, Jianteng Chen, Wenlong Ma, Chenglong Li, Lin Wang, Hengzhen Feng, Lu Shi, et al. Robo-gs: A physics consistent spatial-temporal model for robotic arm with hybrid representation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 15379–15386. IEEE, 2025.

[Ma *et al.*, 2025] Zipei Ma, Junzhe Jiang, Yurui Chen, and Li Zhang. Béziergs: Dynamic urban scene reconstruction with bézier curve gaussian splatting. In *ICCV*, 2025.

[Qureshi *et al.*, 2025] M Nomaan Qureshi, Sparsh Garg, Francisco Yandun, David Held, George Kantor, and Abhisesh Silwal. Splatsim: Zero-shot sim2real transfer of rgb manipulation policies using gaussian splatting. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6502–6509. IEEE, 2025.

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

[Wang *et al.*, 2025] Yifan Wang, Peishan Yang, Zhen Xu, Jiaming Sun, Zhanhua Zhang, Yong Chen, Hujun Bao, Sida Peng, and Xiaowei Zhou. Freetimegs: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21750–21760, 2025.

[Wu *et al.*, 2024] 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.

[Wu *et al.*, 2025] Yuxuan Wu, Lei Pan, Wenhua Wu, Guangming Wang, Yanzi Miao, Fan Xu, and Hesheng Wang. Rl-gsbridge: 3d gaussian splatting based real2sim2real method for robotic manipulation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 192–198. IEEE, 2025.

[Xue *et al.*, 2025] Zhengrong Xue, Shuying Deng, Zhenyang Chen, Yixuan Wang, Zhecheng Yuan, and Huazhe Xu. Demogen: Synthetic demonstration generation for data-efficient visuomotor policy learning. *arXiv preprint arXiv:2502.16932*, 2025.

[Yang *et al.*, 2024] 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.

[Yang *et al.*, 2025] Sizhe Yang, Wenye Yu, Jia Zeng, Jun Lv, Kerui Ren, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Novel demonstration generation with gaussian splatting enables robust one-shot manipulation. *arXiv preprint arXiv:2504.13175*, 2025.

[Zhang *et al.*, 2024] Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pages 199–216. Springer, 2024.

[Zhang *et al.*, 2025a] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In *2025 International Conference on 3D Vision (3DV)*, pages 1091–1102. IEEE, 2025.

[Zhang *et al.*, 2025b] Xinjie Zhang, Zhening Liu, Yifan Zhang, Xingtong Ge, Dailan He, Tongda Xu, Yan Wang, Zehong Lin, Shuicheng Yan, and Jun Zhang. Mega: Memory-efficient 4d gaussian splatting for dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27828–27838, 2025.

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

[Zhu *et al.*, 2024] Ruijie Zhu, Yanzhe Liang, Hanzhi Chang, Jiacheng Deng, Jiahao Lu, Wenfei Yang, Tianzhu Zhang, and Yongdong Zhang. Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37:101790–101817, 2024.