arXiv:2511.14540v2 [cs.CV] 29 Jun 2026

Interaction-Aware 4D Gaussian Splatting for Dynamic Hand-Object Interaction Reconstruction

Hao Tian[1,2,5], Chenyangguang Zhang[3], Rui Liu[4], Wen Shen[1,5], and Xiaolin Qin[1,5]*

[1] Chengdu Institute of Computer Applications, Chinese Academy of Sciences, Chengdu, China
[2] PetroChina (Beijing) Digital Intelligence Research Institute Co., Ltd., Beijing, China
[3] Tsinghua University, Beijing, China
[4] Minzu University of China, Beijing, China
[5] University of Chinese Academy of Sciences, Beijing, China
haotian23ucas.github.io/Interaction-Aware-Gau

Abstract. This paper focuses on a challenging setting of simultaneously modeling geometry and appearance of hand-object interaction scenes without any object priors. We follow the trend of dynamic 3D Gaussian Splatting based methods, and address several significant challenges. To model complex hand-object interaction with mutual occlusion and edge blur, we present interaction-aware hand-object Gaussians with newly introduced optimizable parameters aiming to adopt piecewise linear hypothesis for clearer structural representation. Moreover, considering the complementarity and tightness of hand shape and object shape during interaction dynamics, we incorporate hand information into object deformation field, constructing interaction-aware dynamic fields to model flexible motions. To further address difficulties in the optimization process, we propose a progressive strategy that handles dynamic regions and static background step by step. Correspondingly, explicit regularizations are designed to stabilize the hand-object representations for smooth motion transition, physical interaction reality, and coherent lighting. Experiments show that our approach surpasses existing dynamic 3D-GS-based methods and achieves state-of-the-art performance in reconstructing dynamic hand-object interaction.

1 Introduction

Accurate reconstruction of hand–object interaction (HOI) is vital for VR and robotics [9], requiring precise shape modeling and interaction capture. Despite the apparent simplicity of daily actions like grasping or drinking, they involve complex contact dynamics and severe occlusions that remain challenging to model. Previous works [36, 40] reconstruct HOI scenes by treating interactive objects as known, relying on specific object poses or templates. However, acquiring such poses or templates is costly, limiting their industrial applicability. Existing methods try to reduce the reliance on precise pose estimation and

* Corresponding author.

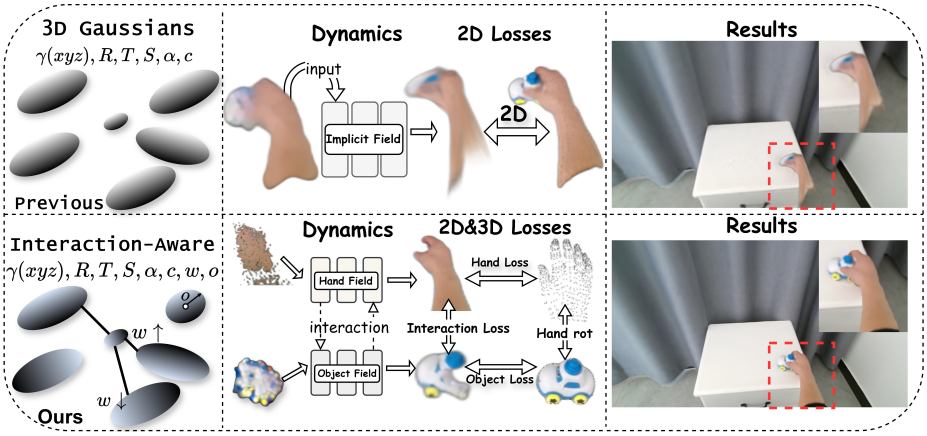

Fig. 1. Differences between traditional 3D Gaussian-based hand-object reconstruction and our interaction-aware modeling. Conventional 3D Gaussian approaches model the entire HOI scene with a single, unified implicit field and rely primarily on 2D supervision. This design often leads to geometric ambiguities during close interactions—such as collapsed clearances, blurred contact boundaries, and non-physical merging of hand and object surfaces (top). In contrast, our method explicitly decouples hand and object representations into separate fields, introduces interaction-aware parameters ($\boldsymbol{w}, \boldsymbol{o}$) to modulate occlusion and edge sharpness, and leverages interaction-aware losses to preserve fine-grained spatial relationships, enabling accurate and disentangled dynamic reconstruction (bottom).

templates by several routes. [34, 45] employs SDF-based approaches, integrating SDF to reconstruct dynamic hand-object scenes using neural networks, without the need for specific object priors. However, these methods focus primarily on geometry reconstruction without appearance. With the advent of NeRF [25], some researchers [21, 55] explore HOI scene reconstruction with both geometry and appearance by training implicit fields. However, due to the inefficiency of backward mapping-based ray-rendering algorithms [25], these methods require significant time and computational resources. Recently, 3D Gaussian Splatting (3D-GS) [16] has demonstrated superior fidelity and speed in static scene reconstruction. Some works [13, 46, 50] attempt dynamic reconstruction using 3D-GS, but they struggle with complex HOI scenarios involving heavy occlusion and irregular rotations, failing to capture accurate interaction dynamics. Although EgoGaussian [54] targets HOI reconstruction, it requires object pose estimation and only presents results for interactive objects without effective hand representation. In this work, we address the limitations of 3D-GS-based methods by proposing a model to simultaneously reconstruct the entire HOI scene without requiring any object priors.

To successfully handle such a practical setting presents significant challenges. First, drastic motions, mutual occlusion and blur during interaction cause misalignment and excessive overlap among Gaussians. To address this, we model the interaction as a piecewise linear process and present a novel representation termed interaction-aware hand-object Gaussians. It introduces two parameters

over the traditional 3D-GS representation: weight w and radius o. The weight w balances motion smoothness and noise reduction, with smaller values indicating weak structural information or occlusion. The radius o controls edge sharpness, where smaller values produce clearer contours. The combination of w and o effectively models the complex dynamic interaction, reducing blurring at interaction boundaries and enhancing visual quality. Second, previous methods [13,50] use a single field to model Gaussian transformations, which is insufficient for capturing drastic and highly localized motions in HOI scenes, often leading to significant loss of fine-grained motion details. On the other hand, simply using separate fields for hand and object deformation overlooks their mutual interaction and geometric coupling during contact. To address this, we incorporate key-frame hand positions into the object field, enabling interaction-aware transformations that accurately capture dynamic changes caused by hand grasping. Third, considering the flexible motions, irregular rotations, and frequent occlusions in HOI scenes, it is difficult to directly utilize traditional 3D-GS optimization [16,46,50] to achieve decent rendering quality. To address this, we design explicit interaction-aware regularizations to explicitly stabilize the position and rotation of hand-object Gaussians. Furthermore, we propose a progressive optimization mechanism to achieve physically realistic hand-object interaction, ensure smooth edge transitions, and enhance coherence in complex HOI scenes.

Our contributions are summarized as follows:

- We propose a novel interaction-aware hand-object Gaussian representation to model HOI scenes without any object priors, effectively addressing mutual occlusion and edge blur during interactions.
- We enhance the object field with hand information to capture interaction-induced deformation.
- We use a progressive optimization strategy with explicit 3D losses to benefit the fitting of the Gaussians during dynamic reconstruction.
- Experiments show that our approach surpasses state-of-the-art baselines, achieving superior performance in reconstructing dynamic HOI scenes.

2 Related Works

Hand Representation. Early approaches [27, 28] focused on estimating 2D or 3D keypoints from images. The introduction of statistical hand models like MANO [35] revolutionized parametric hand representation by jointly encoding pose, shape, and 3D vertices. Recent method [1] typically employs regression networks to predict MANO parameters directly from images and optimize shape parameters for alignment. However, these methods suffer from error propagation: initial MANO inaccuracies accumulate downstream, causing cascading reconstruction errors. To mitigate this, we propose a hand field to decouple hand deformation from strict MANO parameter dependencies. Recent works have extended hand modeling using neural and Gaussian representations [4, 14, 32]. In contrast, our work targets prior-free, monocular, egocentric HOI reconstruction without requiring multi-view inputs or personalized templates.

Hand-Object Reconstruction. Reconstructing hand-object interaction from video remains a significant challenge in computer vision and graphics. Previous works fall into two categories. The first [3, 8, 26, 42] reconstructs hand-object interactions from multiview sources by fitting objects into 2D images using 3D object priors. However, these methods heavily rely on accurate 3D priors, which are costly to obtain. The second stream [6, 11, 36] pre-learns object templates to reduce reliance on priors. For example, the MANO model [35] represents the canonical hand space, with linear blending skinning driving the hand template. EgoGaussian [54] reconstructs egocentric interaction scenes using 3D Gaussian [16] by separating dynamic objects from the static background. However, it is sensitive to object pose, probably failing under inaccurate poses, and excludes interacting hands from reconstruction. As a result, it misses the complete hand–object interaction context. Recent work BIGS [29] reconstructs hand-object interactions from monocular video using 3D-GS and a diffusion prior, but assumes a known object mesh and excludes the background. In contrast, our method is category-agnostic, requires no object priors, and reconstructs the full hand-object interaction scene. A growing line of work leverages text-to-image diffusion priors to hallucinate unseen object geometry under heavy occlusion. Ye *et al*. [53], G-HOP [52], and MagicHOI [43] share our prior-free goal but trade flexibility for diffusion-induced regularization, while our approach reconstructs HOI scenes without any external generative model. For articulated objects with parametric priors, BimArt [58] and SyncDiff [12] target an orthogonal problem space. We position our work as complementary: prior-free reconstruction of egocentric HOI scenes involving rigid objects, without diffusion priors or object templates.

Dynamic Scene Reconstruction. With the advent of NeRF [25], many works [7,19,30,31,33,41,49] use MLPs to represent implicit spaces as deformation fields with temporal information. However, their extensive training time limits practical applicability. 3D Gaussian Splatting (3D-GS) [16] emerges as a promising alternative for scene reconstruction. Methods like [13, 15, 17, 18, 20, 22, 24, 37, 38, 46–48, 50, 57] explore dynamic reconstruction using 3D-GS. For instance, [50] uses MLPs to learn Gaussian position offsets per timestamp, which increases training time. [13] introduces sparse control points to deform Gaussians, but in hand-object interaction (HOI) scenes, redundant points lead to inaccuracies and image tearing, failing to capture intricate interactions. These methods [13, 46, 50] input all Gaussians into a single MLP, which struggles to accurately model complex interactive motions. To overcome significant challenges posed by HOI scenes, our method introduces a novel interaction-aware hand-object Gaussian representation, with adaptive losses and a progressive optimization strategy.

3 Method

Our goal is to reconstruct dynamic hand-object interaction (HOI) scenes from RGB egocentric videos at *arbitrary timestamps* without relying on any object *shape priors*. We utilize three implicit fields to model the dynamic HOI scenes: the hand field $\mathcal{F}_{\mathrm{H}}$ and object field $\mathcal{F}_{\mathrm{O}}$ to approximate the shape of the dynamic

HOI region, as well as the background field $\mathcal{F}_{\text{BG}}$ to create a clean background and facilitate subsequent joint optimization. Separate modeling allows capturing clear hand-object appearance and stable background scene in drastically changing dynamic scenarios. First, by treating hand and object modeling differently, significant occlusions could be solved via more detailed supervision. Second, backgrounds require low-frequency updates, while hand-object interactions demand high-frequency modeling. Meanwhile, collaborating with such dynamic implicit fields, we adaptively improve the Gaussian Splatting representation for HOI scenarios, addressing occlusion and contour clarity issues during the interaction. In optimization, we utilize only the hand's MANO parameters predicted by an off-the-shelf hand tracker [51], which incurs negligible computational overhead compared to Gaussian optimization (typically $< 3\%$ of total runtime). These pose estimates provide coarse 3D hand guidance to accelerate convergence without requiring any object priors. To ensure physically plausible interactions, we further introduce an interaction loss. A progressive and collaborative optimization framework is then devised to achieve high-quality HOI scene reconstruction through this lightweight 3D supervision and our interaction-aware representation.

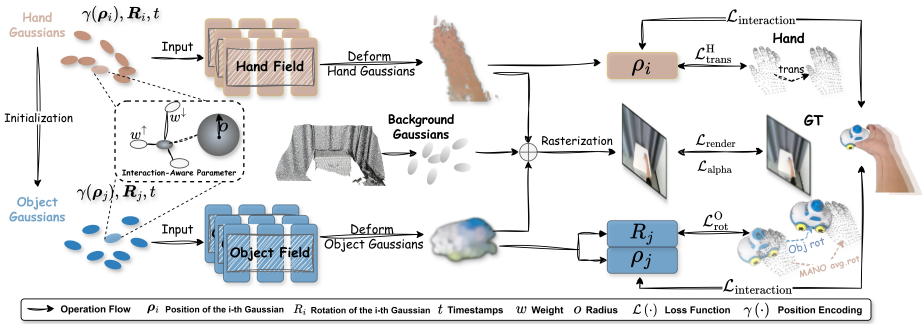

Fig. 2. Overview of interaction-aware hand-object Gaussians. We propose a novel framework for reconstructing dynamic HOI scenes from RGB videos without object shape priors. The framework consists of three components: (1) *Specialized Implicit Fields*: separate hand, object, and background fields disentangle dynamic interactions, with hand/object fields capturing high-frequency deformations and occlusions (leveraging hand information for object's interaction-aware deformation) while the background field maintains low-frequency stability; (2) *Interaction-aware Gaussian*: enhances representation with adaptive weights w and radius o to address contour ambiguity and occlusions; (3) *Progressive Optimization*: combines explicit supervision with physical interaction constraints for efficient convergence.

3.1 Preliminaries: Deformable Gaussian Splatting

3D Gaussian Splatting (3D-GS) [16] represents 3D scene features using five parameters: position, transparency, spherical harmonic coefficients, rotation, and

scaling. 3D-GS explicitly defines each 3D Gaussian ellipsoid in space using a covariance matrix Σ and the position vector $\boldsymbol{\rho}$, as shown in the following equation:

$$\mathcal{G}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma|^{\frac{1}{2}}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\rho})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\rho})), \tag{1}$$

here the Σ matrix can be decomposed into a rotation $\boldsymbol{R}$ and a scaling $\boldsymbol{S}$ by $\Sigma = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{R}^\top$.

Recent works [13,46,50] combine 3D-GS with deformation fields for dynamic scenes, using an MLP to warp points from canonical to target space:

$$\begin{aligned} \mathcal{F}_\theta(\boldsymbol{x}, \boldsymbol{t}) &= (\delta\boldsymbol{x}^t, \delta\boldsymbol{r}^t, \delta\boldsymbol{s}^t), \\ \boldsymbol{x}^t &= \boldsymbol{x} + \delta\boldsymbol{x}^t. \end{aligned} \tag{2}$$

Each initialized 3D Gaussian's center $\boldsymbol{x}$ is input to deformation field $\mathcal{F}_\theta$, which outputs time-dependent offsets ($\delta\boldsymbol{x}^t$, $\delta\boldsymbol{r}^t$, $\delta\boldsymbol{s}^t$) to wrap canonical Gaussians to target space.

3.2 Interaction-Aware Hand-Object Gaussians

To effectively capture the complex spatiotemporal motion in the Hand-Object Interaction (HOI) scene, we propose to decompose the dynamic HOI scene into three sets of Gaussians and model each part individually. Moreover, we improve traditional 4D Gaussian representations to overcome the issues representing complex HOI motions in dynamic scenarios. Traditional 4D Gaussians [13,46,50] tend to neglect mutual influences between different interacted Gaussians. Moreover, it is hard to depict contour edges in interaction process, leading to texture drift and edge blur. Inspired by [13], we introduce two additional learnable parameters weight $\boldsymbol{w} \in \mathbb{R}^+$ and radius $\boldsymbol{o} \in \mathbb{R}^+$, forming a novel representation termed the Interaction-Aware Hand-Object Gaussian $\mathcal{G}_{\text{HO}}$. $\mathcal{G}_{\text{HO}}$ focuses on interaction-aware modeling via: (1) Weight $\boldsymbol{w}$ smooths the motion and reduces noise, it models mutual occlusion during interactions, where a small weight $\boldsymbol{w}$ indicates weak structural information and occlusion; (2) Radius $\boldsymbol{o}$ captures edge details, where a small radius $\boldsymbol{o}$ corresponds to sharper geometric contours near edges; (3) Since $\boldsymbol{w}$ is larger near the current Gaussian and smaller farther from the edge, the combination of weight $\boldsymbol{w}$ and radius $\boldsymbol{o}$ effectively handles edge blurring between hand-object interactions and the background. $\mathcal{G}_{\text{HO}}$ is given as follows:

$$\mathcal{G}_{\text{HO}} = \{\boldsymbol{x}_i\boldsymbol{y}_i\boldsymbol{z}_i, \boldsymbol{R}_i, \boldsymbol{T}_i, \boldsymbol{S}_i, \boldsymbol{\alpha}_i, \boldsymbol{c}_i, \boldsymbol{w}_i, \boldsymbol{o}_i\}. \tag{3}$$

Due to different characteristics of hand motions, object motions and background scenes, we introduce each component's dynamics separately below.

Hand Gaussians. Hand Gaussians $\mathcal{G}_{\text{H}}$ has the same optimizable parameters with $\mathcal{G}_{\text{HO}}$, and is modeled with a hand-implicit field $\mathcal{F}_{\text{H}}$ to capture the time-varying transformation of hand motion. This field $\mathcal{F}_{\text{H}}$ takes the timestamp t and the canonical position ($\boldsymbol{x}_i\boldsymbol{y}_i\boldsymbol{z}_i$) of the i-th hand Gaussian as inputs. Since our setting requires modeling dynamic HOI scenes at any time from any view pose,

we use t as the input and construct the following formula (details provided in the supplementary material):

$$\Delta\mathcal{G}_{\mathrm{H}} = \mathcal{F}_{\mathrm{H}}\left\{\gamma\left(\boldsymbol{x}_i\boldsymbol{y}_i\boldsymbol{z}_i\right), \gamma\left(t\right)\right\}, \tag{4}$$

$\gamma(\cdot)$ is positional encoding [2, 50]. Adding $\mathrm{noise}_{\mathrm{smooth}}$ [50] to $\gamma(t)$ prevents over-smoothing and retains hand details while fitting coarse geometry.

Object Gaussians. Hand-object interactions often cause deformations or occlusions (e.g., holding). To enhance the ability of the object field $\mathcal{F}_{\mathrm{O}}$ to capture interaction-aware deformations, we introduce hand position as an additional input to the object field. This overcomes the limitation of implicit methods [46,50], which generate global offsets without explicit hand-object interaction modeling. The object field $\mathcal{F}_{\mathrm{O}}$ takes both hand and object positions as inputs, formulated as follows (details in supplementary material):

$$\Delta\mathcal{G}_{\mathrm{O}} = \mathcal{F}_{\mathrm{O}}\left\{\gamma\left(\left(\boldsymbol{x}_i^k\boldsymbol{y}_i^k\boldsymbol{z}_i^k\right) \oplus \left(\boldsymbol{x}_j^k\boldsymbol{y}_j^k\boldsymbol{z}_j^k\right)\right), \gamma\left(t\right)\right\}. \tag{5}$$

Here, $\oplus$ concatenates hand $(\boldsymbol{x}_i^k, \boldsymbol{y}_i^k, \boldsymbol{z}_i^k)$ and object $(\boldsymbol{x}_j^k, \boldsymbol{y}_j^k, \boldsymbol{z}_j^k)$ positions with the canonical Gaussian position at key-frame k—the moment just before hand-object interaction. The object-implicit field $\mathcal{F}_{\mathrm{O}}$ predicts time-varying offsets $\Delta\mathcal{G}_{\mathrm{O}}$ at timestamp t, and uses linear annealing of $\mathrm{noise}_{\mathrm{smooth}}$ [50] to stabilize training.

Background Gaussians. We construct background Gaussians to better capture the smoothness and static nature of the background and avoid the unstable dynamic changes of the background Gaussian distribution caused by the interaction of foreground hand-object, which will affect the rendering quality of the background [13, 46]. Background Gaussians $\mathcal{G}_{\mathrm{BG}}$ are based on the Deform3DGS model [50]. Their positions change over time, as formulated in Eq. (2), using the background-implicit field $\mathcal{F}_{\mathrm{BG}}$ with timestamps t.

3.3 Explicit Interaction-Aware Regularizations

2D regularization is to constrain pixel errors in image space. However, this is insufficient due to significant occlusion and drastic motion in hand-object interaction scenes. To enable Gaussians to accurately and efficiently model complex hand-object interactions, besides 2D supervision, we introduce explicit 3D regularizations from interaction priors. Critically, these losses require no object shape or pose priors. We only leverage a lightweight, off-the-shelf hand tracker [51] to provide coarse 3D hand guidance. These include object, hand, and interaction losses to stabilize the rotation and transformation of interaction-aware hand-object Gaussians, ensuring physically plausible dynamics without relying on any object-specific assumptions.

Hand Loss. Hand movement in HOI scenes is fast, making dynamic Gaussian fitting much slower and more challenging. Since MANO vertices explicitly represent the position of each point, we design a hand loss to optimize the translation of hand Gaussians. To track their translation, we use a single Chamfer Distance (CHD) to supervise Gaussian translation in 3D space, we compute its

distance to the nearest vertex on the MANO $\mathcal{V}_h$. This loss measures the distance from each hand Gaussian to its closest point on the MANO vertices, encouraging the Gaussians to populate the hand surface, formulated as follows:

$$\mathcal{L}_{\text{trans}}^{\text{H}} = \frac{1}{N}\sum_{i=1}^{N} \min_{\mathbf{v}\in\mathcal{V}_h} \|(\boldsymbol{x}_i\boldsymbol{y}_i\boldsymbol{z}_i) - (\boldsymbol{x}_v\boldsymbol{y}_v\boldsymbol{z}_v)\|_2^2\,, \tag{6}$$

where $(\boldsymbol{x}_i\boldsymbol{y}_i\boldsymbol{z}_i)$ denotes the i-th Gaussian position, $(\boldsymbol{x}_v\boldsymbol{y}_v\boldsymbol{z}_v)$ represents the filtered points within MANO vertices' range (addressing arm-hand discrepancies).

Object Loss. Since object motion—both translation and rotation—tends to follow hand motion, we regularize the object field using hand cues. While translation is implicitly aligned via spatial proximity, rotation often suffers from non-physical flipping, especially in passive contacts. We thus introduce a hand-guided rotation loss $\mathcal{L}_{\text{rot}}^{\text{O}}$ that aligns object Gaussians with the hand's dominant rotational trend.

In tightly coupled interactions (e.g., grasping), object rotation typically follows the hand's dominant rotational trend. We compute a global prior $\mathbf{R}_{\text{hand}}^{\text{target}}(t) \in \text{SO}(3)$ by averaging relevant MANO joint rotations via SVD-based averaging [10]. To apply regularization only during contact, we modulate the loss with the interaction-aware weight $\boldsymbol{w}_j^{\text{O}}$ of each object Gaussian j, which is small under occlusion/non-contact and large when engaged. The contact-aware weight is:

$$\omega_j(t) = \sigma(\boldsymbol{w}_j^{\text{O}}), \tag{7}$$

where σ is the sigmoid function. The final loss is:

$$\mathcal{L}_{\text{rot}}^{\text{O}} = \mathbb{E}_t\left[\frac{1}{N}\sum_{j=1}^{N}\omega_j(t)\left\|\log\left((\mathbf{R}_{\text{hand}}^{\text{target}}(t))^{-1}\mathbf{R}_{t,j}^{\text{obj}}\right)\right\|^2\right], \tag{8}$$

$\mathbf{R}_{t,j}^{\text{obj}} \in \text{SO}(3)$ is the predicted rotation of the j-th object Gaussian at timestamp t, and $\log(\cdot)$ denotes the logarithmic map from SO(3) to its Lie algebra $\mathfrak{so}(3)$. This metric-aware penalty suppresses implausible rotations while preserving local deformation freedom.

Interaction Loss. Reconstruction of grasping interactions often suffers from edge blurring and mutual occlusion of Gaussians. To regularize the physical reality, we introduce the self-supervised Chamfer distance between hand and object Gaussians. Our approach models the hand and object separately, explicitly defining their positions. This allows us to introduce an interaction loss to ensure proper grasping, formulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{interaction}} =& \frac{1}{\max(|C_H|,\epsilon)}\sum_{i\in C_H}\min_{j\in C_O}\|\mathbf{p}_i - \mathbf{p}_j\|_2^2 \\ &+ \frac{1}{\max(|C_O|,\epsilon)}\sum_{j\in C_O}\min_{i\in C_H}\|\mathbf{p}_i - \mathbf{p}_j\|_2^2,\end{aligned} \tag{9}$$

where $\epsilon = 10^{-6}$ avoids division by zero when no contacts are detected. While this loss promotes hand–object proximity, it does not prevent interpenetration. We therefore use a separate penetration loss (supplementary material) that penalizes overlapping or overly close Gaussians from the hand and object. This loss ensures physical realism and visual quality by minimizing hand–object Gaussian distances while avoiding overlap.

3.4 Progressive Optimization

In the Hand-Object Interaction (HOI) scene, complex rotations, translations, and occlusions are common. Directly optimizing all Gaussians leads to slow convergence and positional misalignment. To address these issues, we propose a progressive optimization strategy for learning individual implicit fields, which operates in five phases as below: initialization, warm-up, HOI refinement, background optimization, and collaborative reconstruction.

Initialization. The MANO vertices [35] provide a coarse initialization for hand geometry, obtained from an off-the-shelf hand tracker [51] and used solely to bootstrap $\mathcal{G}_{\mathrm{H}}$. In contrast, for the object, we do not assume any shape, category, or 3D bounding box prior. Instead, $\mathcal{G}_{\mathrm{O}}$ is randomly initialized by uniformly sampling 3D points within an expanded axis-aligned bounding box (AABB) of the MANO vertices. For the background, $\mathcal{G}_{\mathrm{BG}}$ is initialized from SfM-based sparse point clouds.

Warm-up. During the warm-up phase, we use the proposed 3D losses besides the fundamental 2D losses. For the hand field, we employ $\mathcal{L}_{\mathrm{trans}}^{\mathrm{H}}$ to guide the deformation of hand Gaussians $\mathcal{G}_{\mathrm{H}}$, ensuring alignment with the target pose. For the object field, to stabilize interaction-aware transformations, we use $\mathcal{L}_{\mathrm{rot}}^{\mathrm{O}}$. During the warm-up phase, we periodically apply gradient-based density adjustments [16] to optimize the initial Gaussian distribution.

HOI Refinement. We adaptively refine Gaussians by assigning each Gaussian i a learnable weight $\boldsymbol{w}_i$ and radius $\boldsymbol{o}_i$, $\boldsymbol{o}_i$ controls its local influence range. The final refinement weight for the k-th nearest neighbor of Gaussian i is obtained by: (1) computing spatial proximity weights $\boldsymbol{w}_{ik}^{\mathrm{spatial}}$ for the K nearest neighbors via a RBF kernel on distance $\boldsymbol{d}_{ik}$ and $\boldsymbol{o}_i$ (Eq. 10), (2) normalizing these weights to sum to one (Eq. 11), and (3) modulating them with a global importance weight $\sigma(\boldsymbol{w}_i)$ (Eq. 12). This allows joint learning of global importance $\boldsymbol{w}_i$ and local context.

$$\boldsymbol{w}_{ik}^{\mathrm{spatial}} = \exp\left(-\frac{\boldsymbol{d}_{ik}^2}{2\boldsymbol{o}_i^2}\right), \quad k \in \mathcal{N}_K(i), \tag{10}$$

where $\mathcal{N}_K(i)$ denotes the set of K nearest neighbor Gaussians for the i-th Gaussian, $\boldsymbol{d}_{ik}$ is the Euclidean distance between the centers of Gaussians i and k, and $\boldsymbol{o}_i$ is the learnable radius parameter associated with Gaussian i.

$$\hat{\boldsymbol{w}}_{ik}^{\mathrm{spatial}} = \frac{\boldsymbol{w}_{ik}^{\mathrm{spatial}}}{\sum_{j \in \mathcal{N}_K(i)} \boldsymbol{w}_{ij}^{\mathrm{spatial}}}. \tag{11}$$

The refined weight of Gaussian i's k-th neighbor is:

$$\hat{\boldsymbol{w}}_i^k = \sigma(\boldsymbol{w}_i) \cdot \hat{\boldsymbol{w}}_{ik}^{\text{spatial}}, \quad \text{for} \quad k \in \mathcal{N}_K(i), \tag{12}$$

where $\sigma(\cdot)$ is the sigmoid function ensuring $\boldsymbol{w}_i \in (0, 1)$.

Additionally, we query the hand implicit field $\mathcal{F}_{\text{H}}$ and the object implicit field $\mathcal{F}_{\text{O}}$ to obtain their respective rotation matrices $\left(\Delta\boldsymbol{R}_{6D} \in \mathbb{R}^6\right) \rightarrow \left(\Delta\boldsymbol{R} \in \mathbb{R}^{3\times3}\right)$ (see supplementary material) and translation offset $\Delta(\boldsymbol{x}_k^t \boldsymbol{y}_k^t \boldsymbol{z}_k^t)$. Using a linear blend of local rigid transformations inspired by LBS [39], we refine the pose of the hand-object Gaussians as follows:

$$\begin{aligned} \boldsymbol{T}_k^t &= (\boldsymbol{x}_k^t \boldsymbol{y}_k^t \boldsymbol{z}_k^t) + \Delta(\boldsymbol{x}_k^t \boldsymbol{y}_k^t \boldsymbol{z}_k^t), \\ \boldsymbol{\rho}_i^t &= \sum_{k \in \mathcal{N}_K(i)} \hat{\boldsymbol{w}}_i^k \left(\Delta\boldsymbol{R}_k^t((\boldsymbol{x}_i \boldsymbol{y}_i \boldsymbol{z}_i) - (\boldsymbol{x}_k^t \boldsymbol{y}_k^t \boldsymbol{z}_k^t)) + \boldsymbol{T}_k^t\right). \end{aligned} \tag{13}$$

Here, $(\boldsymbol{x}_i \boldsymbol{y}_i \boldsymbol{z}_i)$ denotes the position of the i-th Gaussian in canonical space, and $\boldsymbol{\rho}_i^t$ represents the deformed position of the i-th Gaussian at timestamp t.

Background Optimization. We pretrain $\mathcal{G}_{\text{BG}}$ for a fixed number of iterations, performing periodic density adjustments [16] to ensure a clean background initialization.

Collaborative Reconstruction. In the final stage, $\mathcal{F}_{\text{H}}$, $\mathcal{F}_{\text{O}}$, and $\mathcal{F}_{\text{BG}}$ independently deform their Gaussians into a shared target space, enabling full HOI scene reconstruction at any timestamp t. Both hand and object Gaussians employ HOI refinement (Eq. (13)) to update their parameters. The optimization is supervised by interaction constraints $\mathcal{L}_{\text{interaction}}$ (Eq. (9)) and 2D regularization terms. This stage ensures physically plausible occlusion relationships, smooth edge transitions, and lighting coherence, improving both the geometric fidelity of reconstructed shapes and the temporal smoothness of their motion dynamics.

4 Experiments

To validate our approach, we conduct comprehensive comparisons with state-of-the-art baselines [13,46,50] for dynamic scene reconstruction on both HOI4D [23] and HO3D [8] datasets. Additionally, we compare with HOLD [5] and BIGS [29], two specialized methods for hand-object interaction reconstruction, on the HO3D dataset. Following [54], we evaluate pure translation and translation-rotation using alternate-frame testing to assess extrapolation to novel interactions. Metrics include PSNR, SSIM [44], and LPIPS [56]. We further perform full-frame evaluation for completeness (Table 1 and 2, Ours*). All experiments run on a single NVIDIA RTX 3090 (24G), achieving optimal performance in 21,000 iterations (1h20m training time).

Implementation Details. We employ K nearest neighbors for refinement and deformation, with the key-frame k set to the timestamp just before hand–object contact. Both Gaussians and the deformation model are optimized using Adam. Hyperparameters, schedules, and auxiliary losses (penetration, momentum, 2D) are provided in the supplementary material. HOI4D [23] provides RGB-D videos

with frame-level hand–object poses and masks; we evaluate on two purely translational and two translation–rotation scenes. HO3D [8] offers real-world 3D pose annotations for actions like pickup and rotation. We use camera 4 from HO3D and select four translation–rotation sequences for egocentric reconstruction.

Methods	**Translation**			**Translation&Rotation**		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
4DGS [46]	24.86-5.46	0.80-0.13	0.47+0.18	*23.68*-0.48	0.85-0.01	0.39+0.02
Deform3DGS [50]	*26.33*-3.99	*0.87*-0.06	**0.29** 0.00	23.57-0.59	**0.89**+0.03	**0.28**-0.09
SC-GS [13]	25.08-5.24	0.84-0.09	*0.46*+0.17	17.32-6.84	0.71-0.15	0.48+0.11
Ours	**30.32**	**0.93**	**0.29**	**24.16**	*0.86*	*0.37*
Ours*	33.03+2.71	0.95+0.02	0.27-0.02	24.02-0.14	0.85-0.01	0.39+0.02

Tab. 1. Quantitative comparison on HOI4D [23]. Best and second-best results are **bolded** and *italicized*, respectively. Differences (red, small font) are relative to the "Ours" row. Ours* denotes full-frame evaluation.

Methods	**Translation&Rotation**		
	PSNR ↑	SSIM ↑	LPIPS ↓
BIGS [29]	3.85-21.34	0.24-0.65	0.70+0.55
HOLD [5]	18.03-7.16	*0.84*-0.05	0.26+0.11
4DGS [46]	19.44-5.75	0.82-0.07	*0.25*+0.10
Deform3DGS [50]	9.68-15.51	0.36-0.53	0.65+0.50
SC-GS [13]	*20.37*-4.82	0.80-0.09	0.26+0.11
Ours	**25.19**	**0.89**	**0.15**
Ours*	25.17-0.02	0.89 0.00	0.16+0.01
BIGS†	24.51-0.68	0.92+0.03	0.07+0.08
Ours†	28.16+2.97	0.95+0.06	0.07-0.08

Tab. 2. Quantitative comparison on HO3D [8]. Best and second-best results are **bolded** and *italicized*. Differences (red font) are relative to "Ours". Ours* denotes full-frame evaluation. † denotes evaluation on hand and object regions only.

4.1 Quantitative Comparisons

HOI4D Dataset. We compare against 4DGS [46], Deform3DGS [50], and SC-GS [13] using official code and original HOI4D resolution (Table 1). 4DGS is sensitive to initialization and underperforms in HOI settings. Deform3DGS and SC-GS, relying on a single deformation field, fail under occlusion and fast motion. Our interaction-aware, progressively optimized model overcomes these issues, reducing occlusion artifacts and blur while preserving hand-object geometry. We achieve a +9% PSNR gain in translation scenes and improve rotation-heavy scene PSNR from 23.57 dB (Deform3DGS) to 24.16 dB.

HO3D Datasets. We downsample all input frames to half resolution for efficient

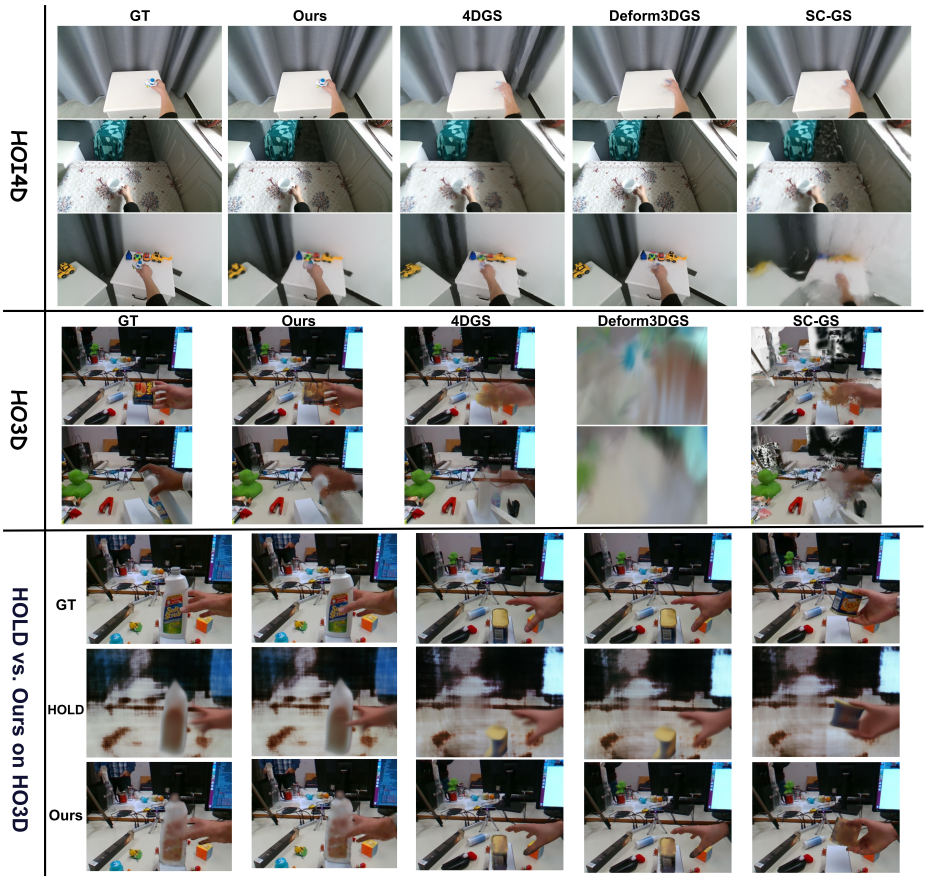

Fig. 3. Qualitative comparison of our approach and the baseline methods. We present reconstructions from our model and SOTA baselines (4DGS [46], Deform3DGS [50], SC-GS [13]) on HOI4D and HO3D datasets.

processing of large-scale sequences. Although HO3D [8] provides camera parameters and hand-object pose estimates, inherent inaccuracies and pose errors adversely affect the performance of all methods (Table 2). 4DGS is highly sensitive to input noise; SC-GS's sparse control points fail to model background–foreground interactions; and Deform3DGS suffers most due to HO3D's pose errors (Appendix D of [50]), causing non-convergence. HOLD [5], designed for geometry rather than view synthesis, lags behind 3DGS-based methods. BIGS [29] reports poor metrics because it reconstructs only foreground hand-object without the background (foreground-only results: † in Table 2). Our approach outperforms all baselines in full-scene reconstruction.

4.2 Qualitative Comparisons

As shown in Fig. 3, our approach surpasses 4DGS [46], Deform3DGS [50], SC-GS [13] and HOLD [5] in both appearance and shape. In the HOI4D scene, base-

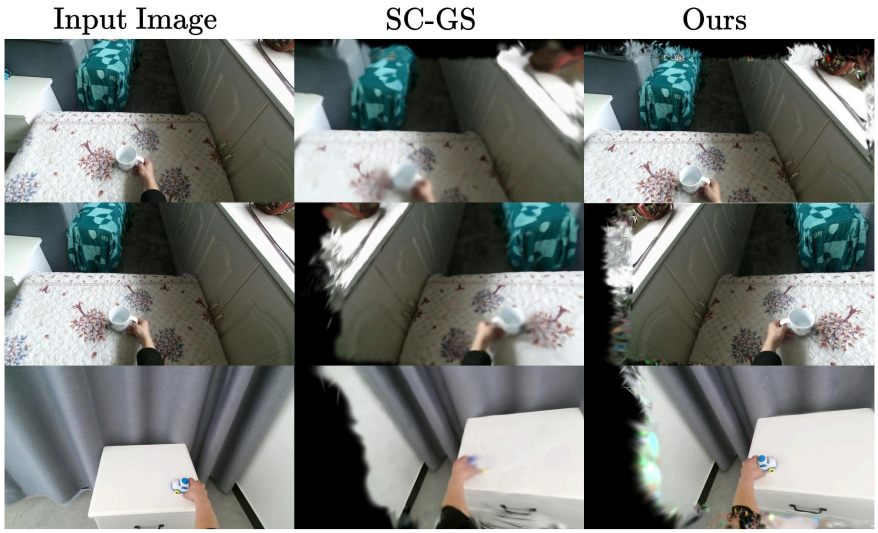

Fig. 4. Extrapolated novel view synthesis of our approach and SC-GS [13]. Our method shows cleaner renderings from novel viewpoints with considerable change, whereas SC-GS suffers from noticeable artifacts.

lines fail to handle Gaussian offsets under dynamic lighting and interaction, while our interaction-aware representation preserves shadows and shapes. Isolated background Gaussians improve contrast and dark details. For the HOI4D scene, separate hand-object modeling and 3D losses effectively constrain interaction-aware deformations, with collaborative reconstruction smoothing motion and occlusion. In the HO3D scene featuring irregular flipping, rotation, and finger flexibility—4DGS falters under noisy or inaccurate data and complex motion. SC-GS loses fine details in interaction zones due to sparse control points and handles occlusions poorly. Deformable3DGS, sensitive to pose errors (Appendix D of [50]), fails to converge on HO3D as errors amplify. HOLD reconstructs hand and object geometry but produces low-quality full-scene renderings. Our method uses w and o to reduce occlusion and blur, achieving superior rendering quality on HO3D and HOI4D.

4.3 Ablation Study

Table 3 reports ablation studies on HOI4D-Scene 1, evaluating the removal of HOI refinement, and the interaction-aware losses and module. We also evaluate robustness to imperfect initialization by adding Gaussian noise $\mathcal{N}(0, \sigma^2)$ to the randomly sampled initial object positions with $\sigma = 0.01$ and 0.05. Table 3 and Fig. 5, our full model achieves near noise-free rendering quality under different noise levels. Removing HOI refinement degrades PSNR/SSIM/LPIPS by 2.2%/1.1%/11.4%; ablating object, hand, or interaction losses causes drops of (4.6%, 1.1%, 8.6%), (1.5%, —, 5.7%), and (3.5%, 1.1%, 14.3%), respectively. To

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o Interaction-Aware Module	28.76-4.20	0.91-0.04	0.40+0.05
w/o HOI Refinement	32.23-0.73	0.94-0.01	0.39+0.04
w/o Object Loss	31.45-1.51	0.94-0.01	0.38+0.03
w/o Hand Loss	32.45-0.51	0.95 0.00	0.37+0.02
w/o Interaction Loss	31.79-1.17	0.94-0.01	0.40+0.05
w/ noise $\sigma = 0.01$	32.80-0.16	0.95 0.00	0.35 0.00
w/ noise $\sigma = 0.05$	32.72-0.24	0.95 0.00	0.35 0.00
Full Model	**32.96**	**0.95**	**0.35**

Tab. 3. Ablation studies on HOI4D. All metrics are reported relative to the Full Model (differences in red, small font). Noise robustness results correspond to Fig. 5.

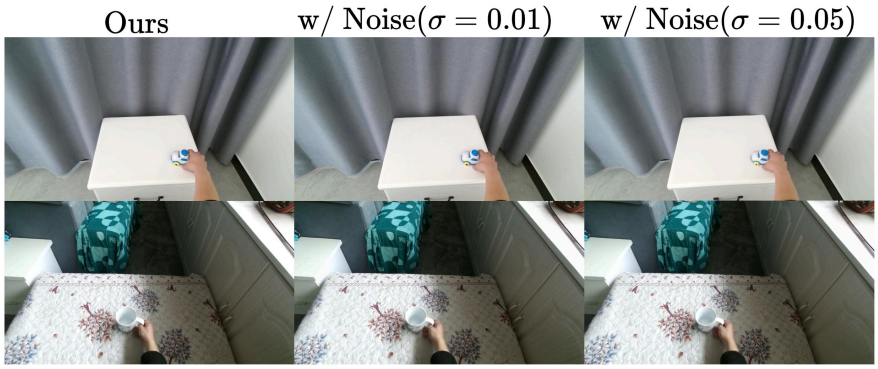

Fig. 5. Our method maintains consistently high rendering quality across different noise levels, showing strong robustness to initialization errors.

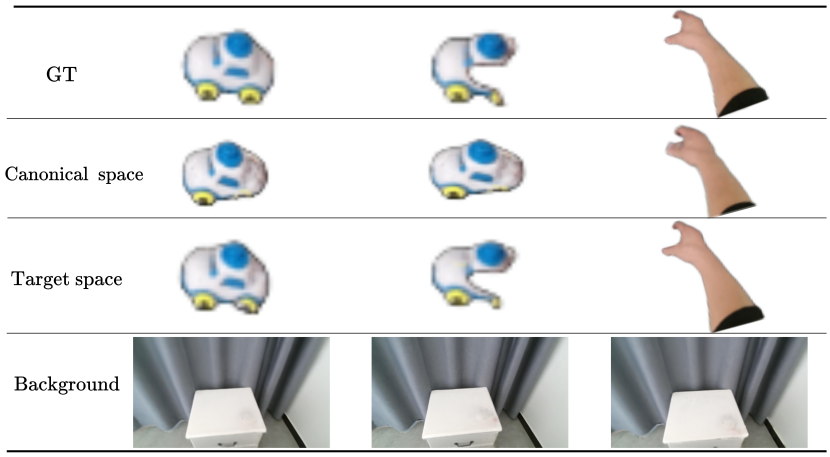

Fig. 6. Disentangled rendering of hand, object, and background. Our method reconstructs each component with high fidelity while maintaining coherent interaction.

validate our interaction-aware design, we conducted an ablation study by eliminating both the field parameters and their associated training scheme from our framework. Tab. 3 clearly shows the performance drop when the interaction-aware module is absent.

5 Conclusion

In this paper, we propose interaction-aware hand-object Gaussians with novel optimizable parameters, adopting piecewise linear hypothesis for a clearer structural representation. This approach effectively captures complex hand-object interactions, including mutual occlusion and edge blur. Leveraging the complementarity and tight coupling of hand and object shapes, we integrate hand information into the object deformation field, constructing interaction-aware dynamic fields for flexible motion modeling. To improve optimization, we propose a progressive strategy that separately handles dynamic regions and static backgrounds. Additionally, explicit interaction-aware regularizations enhance motion smoothness, physical plausibility, and lighting coherence. Experiments show that our approach outperforms the baseline methods, achieving state-of-the-art results in reconstructing dynamic hand-object interactions.

Limitations. As designed for interaction modeling, the workflow consists of progressive optimization stages, which could be unified upon the emergence of new stronger optimizer. Our method struggles with extreme cases (exceedingly rapid motion/complex trajectories, see Supplement), potentially addressable by integrating more interaction priors.

Acknowledgements

This research was partly supported by the Sichuan Science and Technology Program (2025JDDQ0008, 2024NSFJQ0035), and the Talents by Sichuan provincial Party Committee Organization Department.

References

1. Baek, S., Kim, K.I., Kim, T.K.: Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1067–1076 (2019)
2. Bhatnagar, B.L., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. Advances in Neural Information Processing Systems **33**, 12909–12922 (2020)
3. Chen, X., Liu, Y., Ma, C., Chang, J., Wang, H., Chen, T., Guo, X., Wan, P., Zheng, W.: Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13274–13283 (2021)

4. Cong, X., Xing, A., Pokhariya, C., Fu, R., Sridhar, S.: Dytact: Capturing dynamic contacts in hand-object manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops. pp. 7250–7262 (October 2025)
5. Fan, Z., Parelli, M., Kadoglou, M.E., Chen, X., Kocabas, M., Black, M.J., Hilliges, O.: Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 494–504 (2024)
6. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12943–12954 (2023)
7. Guo, X., Sun, J., Dai, Y., Chen, G., Ye, X., Tan, X., Ding, E., Zhang, Y., Wang, J.: Forward flow for novel view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16022–16033 (2023)
8. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3196–3206 (2020)
9. Handa, A., Van Wyk, K., Yang, W., Liang, J., Chao, Y.W., Wan, Q., Birchfield, S., Ratliff, N., Fox, D.: Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system. In: 2020 IEEE International Conference on Robotics and Automation (ICRA). pp. 9164–9170. IEEE (2020)
10. Hartley, R., Trumpf, J., Dai, Y., Li, H.: Rotation averaging. International Journal of Computer Vision **103**(3), 267–305 (2013)
11. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11807–11816 (2019)
12. He, W., Liu, Y., Liu, R., Yi, L.: Syncdiff: Synchronized motion diffusion for multi-body human-object interaction synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 11731–11743 (October 2025)
13. 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4220–4230 (2024)
14. Karunratanakul, K., Prokudin, S., Hilliges, O., Tang, S.: Harp: Personalized hand reconstruction from a monocular rgb video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12802–12813 (June 2023)
15. Katsumata, K., Vo, D.M., Nakayama, H.: A compact dynamic 3d gaussian representation for real-time dynamic view synthesis. In: European Conference on Computer Vision. pp. 394–412. Springer (2024)
16. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
17. Kratimenos, A., Lei, J., Daniilidis, K.: Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In: European Conference on Computer Vision. pp. 252–269. Springer (2024)
18. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6165–6177 (2025)

19. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
20. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21136–21145 (2024)
21. Liu, J.W., Cao, Y.P., Yang, T., Xu, Z., Keppo, J., Shan, Y., Qie, X., Shou, M.Z.: Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18483–18494 (2023)
22. Liu, X., Ren, P., Qi, Q., Sun, H., Zhuang, Z., Wang, J., Liao, J., Wang, J.: Generalizable hand-object modeling from monocular RGB images via 3d gaussians. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), `https://openreview.net/forum?id=JTF0HYfGiL`
23. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21013–21022 (2022)
24. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 2024 International Conference on 3D Vision (3DV). pp. 800–809. IEEE (2024)
25. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
26. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16. pp. 548–564. Springer (2020)
27. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: Ganerated hands for real-time 3d hand tracking from monocular rgb. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 49–59 (2018)
28. Mueller, F., Davis, M., Bernard, F., Sotnychenko, O., Verschoor, M., Otaduy, M.A., Casas, D., Theobalt, C.: Real-time pose and shape reconstruction of two interacting hands with a single depth camera. ACM Transactions on Graphics (ToG) **38**(4), 1–13 (2019)
29. On, J., Gwak, K., Kang, G., Cha, J., Hwang, S., Hwang, H., Baek, S.: Bigs: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 17437–17447 (June 2025)
30. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
31. 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. arXiv preprint arXiv:2106.13228 (2021)
32. Pokhariya, C., Shah, I.N., Xing, A., Li, Z., Chen, K., Sharma, A., Sridhar, S.: Manus: Markerless grasp capture using articulated 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2197–2208 (2024)

33. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
34. Qi, H., Zhao, C., Salzmann, M., Mathis, A.: Hoisdf: Constraining 3d hand-object pose estimation with global signed distance fields. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10392–10402. IEEE (2024)
35. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
36. Rong, Y., Shiratori, T., Joo, H.: Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. arXiv preprint arXiv:2008.08324 (2020)
37. Song, R., Liang, C., Xia, Y., Zimmer, W., Cao, H., Caesar, H., Festag, A., Knoll, A.: Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28031–28041 (2025)
38. Stearns, C., Harley, A., Uy, M., Dubost, F., Tombari, F., Wetzstein, G., Guibas, L.: Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
39. Sumner, R.W., Schmid, J., Pauly, M.: Embedded deformation for shape manipulation. In: ACM siggraph 2007 papers. pp. 80–es (2007)
40. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. pp. 581–600. Springer (2020)
41. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12959–12970 (2021)
42. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision **118**, 172–193 (2016)
43. Wang, S., He, H., Parelli, M., Gebhardt, C., Fan, Z., Song, J.: Magichoi: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5957–5968 (October 2025)
44. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
45. Wen, B., Tremblay, J., Blukis, V., Tyree, S., Müller, T., Evans, A., Fox, D., Kautz, J., Birchfield, S.: Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 606–617 (2023)
46. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
47. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: PhysGaussian: Physics-Integrated 3D Gaussians for Generative Dynamics . In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4389–4398.

IEEE Computer Society, Los Alamitos, CA, USA (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.00420, https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00420

48. Xu, J., Fan, Z., Yang, J., Xie, J.: Grid4d: 4d decomposed hash encoding for high-fidelity dynamic gaussian splatting. Advances in Neural Information Processing Systems **37**, 123787–123811 (2024)
49. Yan, Z., Li, C., Lee, G.H.: Nerf-ds: Neural radiance fields for dynamic specular objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8285–8295 (2023)
50. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20331–20341 (2024)
51. Ye, Y., Feng, Y., Taheri, O., Feng, H., Tulsiani, S., Black, M.J.: Predicting 4d hand trajectory from monocular videos. arXiv preprint arXiv:2501.08329 (2025)
52. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1911–1920 (June 2024)
53. Ye, Y., Hebbar, P., Gupta, A., Tulsiani, S.: Diffusion-guided reconstruction of everyday hand-object interaction clips. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 19717–19728 (October 2023)
54. Zhang, D., Li, G., Li, J., Bressieux, M., Hilliges, O., Pollefeys, M., Gool, L.V., Wang, X.: Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting (2024), https://arxiv.org/abs/2406.19811
55. Zhang, J., Luo, H., Yang, H., Xu, X., Wu, Q., Shi, Y., Yu, J., Xu, L., Wang, J.: Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8834–8845 (2023)
56. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
57. Zhang, S., Wu, G., Xie, Z., Wang, X., Feng, B., Liu, W.: Dynamic 2d gaussians: Geometrically accurate radiance fields for dynamic objects. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 8144–8153 (2025)
58. Zhang, W., Dabral, R., Golyanik, V., Choutas, V., Alvarado, E., Beeler, T., Habermann, M., Theobalt, C.: Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27694–27705 (June 2025)