Grasp in Gaussians: Fast Monocular Reconstruction of Dynamic Hand–Object Interactions

Ayce Idil Aytekin[1,2], Xu Chen[3], Zhengyang Shen[3], Thabo Beeler[3], Helge Rhodin[1,2,4], Rishabh Dabral[1,2], and Christian Theobalt[1,2]

[1] Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
[2] Saarbrücken Research Center for Visual Computing, Interaction and AI
[3] Google
[4] Bielefeld University, Bielefeld, Germany
https://aidilayce.github.io/GraG-page/

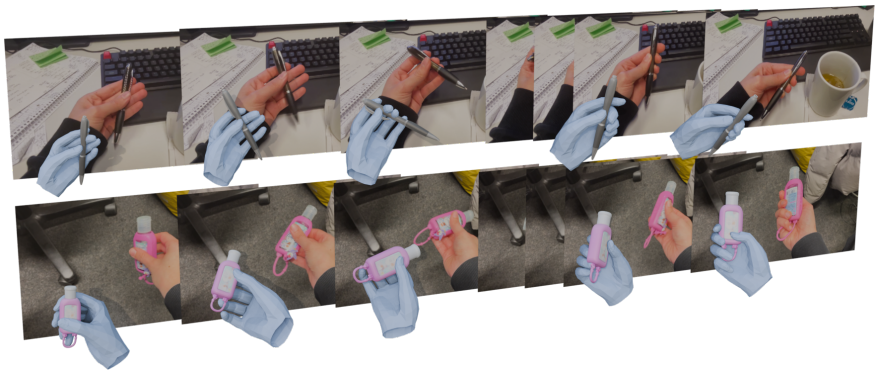

Fig. 1: Grasp in Gaussians (GraG): Given a single monocular video of a hand interacting with an object, *GraG* reconstructs 3D geometry and pose of the hand and the object. Our method is designed to be efficient for long sequences, and can reconstruct in-the-wild captured examples.

Abstract. We present **Grasp in Gaussians (GraG)**, a fast and robust method for reconstructing dynamic 3D hand–object interactions from a single monocular video. Unlike recent approaches that optimize heavy neural representations, our method focuses on tracking the hand and the object efficiently, once initialized from pretrained large models. Our key insight is that accurate and temporally stable hand–object motion can be recovered using a compact Sum-of-Gaussians (SoG) representation, revived from classical tracking literature and integrated with generative Gaussian-based initializations. We initialize object pose and geometry using a video-adapted SAM3D pipeline, then convert the resulting dense Gaussian representation into a lightweight SoG via subsampling. This compact representation enables efficient and fast tracking while preserving geometric fidelity. For the hand, we adopt a complementary strategy: starting from off-the-shelf monocular hand pose initialization, we refine hand motion using simple yet effective 2D joint

and depth alignment losses, avoiding per-frame refinement of a detailed 3D hand appearance model while maintaining stable articulation. Extensive experiments on public benchmarks demonstrate that GraG reconstructs temporally coherent hand-object interactions on long sequences 6.4× faster than prior work while improving object reconstruction by 13.4% and reducing hand's per-joint position error by over 65%.

1 Introduction

Our hands enable us to perform countless interactions with the physical world, from grasping everyday objects to manipulating tools with precision. Reconstructing such dynamic 3D interactions is an important 3D modeling task, with crucial applications in robotics and augmented reality. Our goal is to reconstruct hand-object interaction (HOI), which is the 3D geometry and pose of the hand and the object, from a monocular video. Performing such reconstruction is challenging, as heavy occlusions and depth ambiguity make it a fundamentally ill-posed problem. The challenge is further exacerbated when runtime considerations are introduced.

In the absence of clear visual cues, it becomes important to introduce appropriate priors to curb the ill-posedness of the task. However, past methods have done so with a variety of strong assumptions that often limit their practical applicability to open-world and unconstrained scenarios with complex backgrounds.For instance, many methods require a pre-scanned 3D object template [15, 19, 20, 57], making them unable to reconstruct novel unseen objects. While some template-free methods exist, they are often trained on datasets with a limited number of object instances [21, 23, 59] or are restricted to specific, predefined object categories [60], leading to poor generalization and inaccurate geometry for out-of-distribution objects. Another line of work, in-hand object scanning [16, 22, 63], can reconstruct novel objects but typically assumes a rigid hand pose, failing to capture dynamic, dexterous articulations. Recent video-based approaches such as HOLD [14] and BIGS [34] enable category generalization by optimizing implicit surfaces or Gaussian representations. However, optimizing such volumetric representation across several frames is painstakingly slow (∼10 hours for HOLD, ∼4 hours for BIGS for a video with 100 frames), thereby limiting their use on long, in-the-wild video reconstruction.

To overcome these restrictions, we examine what modern priors can offer and what fast tracking requires. Recent image-to-3D models can produce plausible object geometry from a single view, but naïvely re-running such models frame-by-frame yields jittery poses and inaccurate shape under occlusion. In comparison, classical model-based human/object tracking formulations are fast and stable, but have historically required carefully engineered templates or multi-view capture [47,64], which cannot directly generalize to hand-object reconstruction. In this paper, we show that we can combine the strengths of both worlds by: *using a generative 3D prior to obtain a canonical object, then performing lightweight tracking in a compact representation.*

Our method, dubbed as *Grasp in Gaussians (GraG)*, operates in three stages geared towards balancing computational speed and accuracy. First, we select a small set of informative and diverse keyframes from the input video and reconstruct a canonical object using MV-SAM3D [26], a multi-view extension of SAM3D [6]. MV-SAM3D represents the object as a set of canonical shape tokens (shared across views), which we decode once to obtain a dense 3D Gaussian representation of the object. Second, we extend SAM3D to videos by freezing the canonical shape and estimating per-frame object pose (rotation, translation, and scale) with temporal guidance, which suppresses jitter and prevents occlusion-induced drift. While dense 3D Gaussians are expressive, directly optimizing a high-resolution set of Gaussians across all frames is unnecessary for tracking and becomes computationally expensive. Therefore, we sparsify the decoded dense Gaussians into a compact Sum-of-Gaussians (SoG) model [49] and optimize a differentiable alignment objective that matches its 2D projection to an image SoG constructed from each frame via quad-tree color clustering. For the hand, we refine a Dyn-HaMR [61] initialization using 2D joint reprojection and pointmap depth priors, achieving stable articulation without heavy optimization.

GraG adopts the reconstruction paradigm of having a strong initial prior followed by efficient tracking without additional training. Such a design makes our method particularly well-suited for long, unconstrained videos. Experimental results confirm that GraG achieves state-of-the-art performance on standard benchmarks and generalizes robustly to in-the-wild videos with unseen objects and diverse scenes (*c.f.* Fig. 1). Our method is **highly efficient**: We reduce the runtime of prior works from $\sim$ 3-10 hours to $\sim$30 minutes on long sequences. Our contributions are threefold:

- We propose a fast monocular HOI reconstruction method designed to scale to long, unconstrained videos, drastically reducing the runtime.
- We extend SAM3D to videos by freezing the canonical shape and tracking only per-frame pose with temporal guidance, improving stability under occlusion.
- We introduce compact occlusion-aware SoG tracking for Gaussian-based object assets by sparsifying dense Gaussians and aligning projected object SoG to an image SoG built via quad-tree clustering.

We will release our code for further research upon acceptance.

2 Related Work

3D asset generation: Reconstructing 3D objects from images is a core problem in computer vision. While traditional multi-view methods like Structure from Motion (SfM) [18,45,50,53] are effective, their performance degrades significantly in videos with heavy occlusion and limited viewpoints. To address these limitations, recent work has shifted towards learning-based, single-image 3D generation, trained on large-scale datasets like Objaverse [12], Objaverse-XL [11], and ABO [7]. Diffusion models, in particular, have emerged as the state-of-the-art for generating high-fidelity 3D assets from a single 2D image [6,48,55,56,58]. For

instance, TRELLIS [56] introduces a scalable approach to 3D generation by integrating sparse 3D grids with dense multi-view visual features, enabling versatile decoding into formats like Radiance Fields, 3D Gaussians, and meshes. Recently, building upon TRELLIS, SAM3D [6] proposes pose estimation together with 3D asset generation by separating pose and shape space of the object in the image, resulting in high-quality posed 3D objects. To obtain a strong geometric prior for the object, we employ the multi-view extension of SAM3D.

Gaussian tracking: Gaussian primitives have been used for human [13, 33, 35, 39, 49], hand [3, 46], and object tracking [38] well before Gaussian Splatting (GS) [24]. Beyond their smoothness, SoG [49] exploits a key analytical property: modeling both the image and 3D template with Gaussians makes matching efficient via closed-form Gaussian overlap. Subsequent approaches replaced the resulting sum with more accurate occlusion handling through ray-tracing [39,41], countour-tracking [40], and alpha-blending [24]. To address persistent occlusions and complex topology changes, MVTracker [37] predicts the visibility of individual points within a fused 3D feature point cloud. Similarly, GauSTAR [62] introduces dynamic surface tracking to handle newly appearing geometry. However, these approaches typically necessitate a multi-view setup and rely on high-quality depth maps, either from sensors or external estimators, to maintain 3D consistency. To compensate for large deformations in articulated hand and body motions, structural priors such as parametric model poses are widely integrated into Gaussian-based tracking frameworks. Leveraging the differentiable nature of GS, these methods [9,30] can jointly optimize both hand-object poses and surface geometry through a unified rendering pipeline. Opposed to increasing photorealism with GS and its recent extensions, we revisit the SoG representation, as it alleviates sorting Gaussians back-to-front (for alpha blending) and reduces the number of computations with the sparse Gaussian image representation.

Hand-object reconstruction: Reconstructing HOI has seen a significant shift from template-based methods [4, 10, 29, 52, 57], which are limited to known 3D object models, to more generalizable, template-free techniques. For single-image-to-HOI, recent work has begun to build pipelines that leverage a series of foundation models. In particular, EasyHOI [31] and FollowMyHold [1] utilize separate models for segmentation, inpainting, and initial 3D generation to achieve impressive generalization to in-the-wild images. However, as single-frame methods, they cannot leverage the rich temporal cues available in videos. For video-based reconstruction, HOLD [14] achieves category-agnostic results by optimizing a neural implicit field per-scene, but this process is slow and fails to reconstruct geometry for consistently occluded regions. To address occlusion, subsequent works like BIGS [34] and MagicHOI [54] incorporate generative priors, using Score Distillation Sampling (SDS) [36] or Novel View Synthesis (NVS) [28] to complete unseen object parts. While this direction improves completeness, it often comes with substantial per-scene optimization cost and remains sensitive to the quality of the initialization pipeline (commonly SfM-style bootstrapping [43,44]). Consequently, scaling these methods to long, unconstrained interaction videos is challenging: runtime grows quickly with sequence length, and drift can accumulate

under fast motion and hand occlusion. In contrast, GraG shifts computation to a canonical object reconstruction step using a small set of diverse keyframes, and then performs lightweight per-frame tracking in a compact Sum-of-Gaussians formulation with explicit visibility reasoning from the hand mask, enabling fast and stable HOI reconstruction in the wild.

3 Background: SAM3D

Our approach to initialize object poses is built upon SAM3D [6], a recent generative 3D reconstruction model capable of generating an object's geometry and texture, given a single RGB image I and optionally a pointmap P of the scene. Unlike previous single-view reconstruction methods that estimate the shape of the object in an implicit canonical orientation [25, 56], SAM3D explicitly separates the canonical *shape* from the camera-space *layout*.

Concretely, it models 3D reconstruction as a conditional distribution over shape S, texture F, and layout (R, τ, s) as $p(S, F, R, \tau, s \mid I, M, P)$ and learns a generative model to approximate it. Here, R, τ and s denote the object's orientation, translation, and scale, respectively. The model is conditioned on the image I, object mask M, and scene pointmap P. Internally, SAM3D uses a structured multi-modal architecture in which the canonical shape is represented by a set of shape tokens, while the layout parameters are represented by separate low-dimensional tokens. To map these tokens into a high-dimensional feature space, modality-specific input and output projection layers are employed. We denote these tokens by Z^S and Z^P where Z^S encodes the canonical shape and Z^P encodes layout. Z^S can be later decoded to 3D Gaussians [24] and meshes.

4 Method

Given a monocular RGB video $\mathbf{I} = \{\mathbf{I}_t\}_{t=1}^N$ of a hand interacting with a rigid object, our goal is to recover temporally coherent 3D geometry and motion of the hand and the object. This task is highly underconstrained due to depth ambiguity and hand-object occlusions. To tackle it, we leverage the strong data-driven priors provided by generative 3D models like SAM3D and integrate them into the highly efficient classical tracking approach based on Sum-of-Gaussians.

Fig. 2 illustrates the overall schema of our method. This design is motivated by the observation that heavy per-frame 3D reconstruction is unnecessary once a strong canonical prior is available, whereas lightweight tracking can robustly propagate motion through ambiguous and occluded frames.

As preprocessing, we extract hand and object masks M_t^h and M_t^o using SAM3 [5] with a large VLM [51], estimate camera intrinsics/extrinsics (K_t, T_t) and pointmaps P_t with DepthAnything3 [27], initialize the hand trajectory with Dyn-HaMR [61], a MANO-based [42] monocular hand pose estimator, and obtain per-frame hand–object contact flags $c^{ho} \in \{0, 1\}$ (whether the hand is grasping the object) using Gemini 3 [8]. Throughout our discussion, we use superscripts o and h to denote object and hand related variables, respectively.

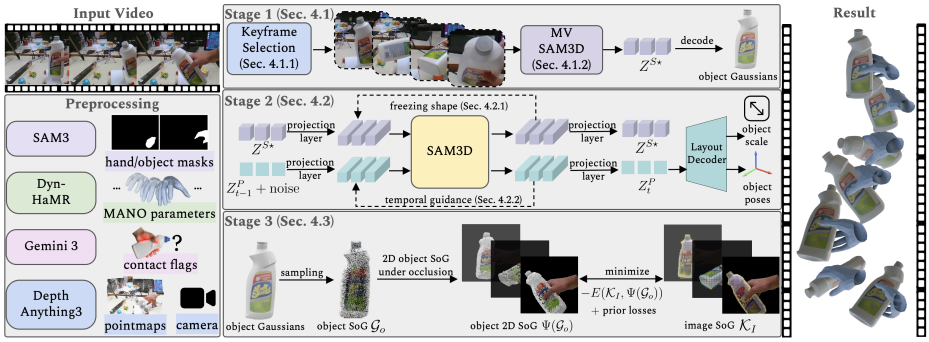

Fig. 2: Overview. Given a monocular video, we recover per-frame hand-object poses and geometry. We first preprocess the video to obtain masks, an initial hand trajectory, per-frame hand–object contact flags, pointmaps, and camera intrinsics/extrinsics. Stage 1 reconstructs a canonical object with MV-SAM3D by selecting keyframes and decoding shape tokens into a dense Gaussian asset (Sec. 4.1). Stage 2 estimates per-frame object pose (and scale) by adapting SAM3D to videos with a frozen canonical shape and temporal guidance (Sec. 4.2). Stage 3 refines hand and object motion using an explicit tracking objective: we sparsify the dense Gaussians into a compact SoG and perform occlusion-aware SoG alignment with lightweight priors (Sec. 4.3).

4.1 Stage 1: Keyframe Selection and Object Reconstruction

Since our system's input is a video, multiple views of the same object are generally available over time, but occluded by the hand. Compared to single-frame reconstruction, these views provide complementary visual cues that can improve object shape estimation. To use this information, and be efficient, we select a compact set of K keyframes $\mathcal{V} = \{t_k\}_{k=1}^{K}$ that are informative and diverse. These video keyframes are treated as multi-view inputs and fed to MV-SAM3D [26], a multi-view extension of the state-of-the-art image-to-3D model SAM3D [6], yielding canonical object shape tokens and a dense 3D Gaussian asset.

Keyframe selection Let $f_t \in \mathbb{R}^d$ be a normalized global descriptor of the object-crop at frame t (extracted from the class-token features from the vision transformer of Depth Anything 3 [27]). With the intuition that a good set of keyframes should be diverse yet balanced, we compute a similarity matrix $S_{ij} = \langle f_i, f_j \rangle$ and select keyframes by minimizing a balance-and-diversity objective:

$$\mathcal{V} = \arg\min_{|\mathcal{V}|=K} \sum_{t\in\mathcal{V}} \left|\bar{S}_t - \tfrac{1}{2}\right| + \left|n_t - \tfrac{1}{2}\right| + \left|\nu_t - \tfrac{1}{2}\right| + \lambda_{\text{div}} \sum_{t\in\mathcal{V}} \max_{s\in\mathcal{V}, s\neq t} S_{ts}, \tag{1}$$

where $\bar{S}_t$ is the mean cosine similarity of frame t to all others, $n_t = \|f_t\|$ is the feature magnitude, and ν_t is the variance across feature dimensions. All $\bar{S}_t$, n_t, and ν_t are min-max normalized. λ_{div} denotes a scalar weighting factor. We solve

Eq. (1) greedily: start from the most "balanced" frame, then iteratively add the next K frames that best trade balance and diversity. Example keyframes can be seen in top left part of the Fig. 2.

Object reconstruction The keyframes $\mathcal{V}$ selected above are then used as input to MV-SAM3D [26]. This simple fusion works well in practice because the canonical shape tokens live in a common object-centric frame [6]. We treat the shape tokens $Z^{S\star}$ from MV-SAM3D as a canonical object representation that can be decoded into Gaussians that we will later sparsify for fast tracking (Sec. 4.3).

4.2 Stage 2: Temporally Stable Object Pose Estimation

In Stage 2, we estimate per-frame object layout (R_t^o, τ_t^o, s^o) while forcing the shape to remain constant over time. This is crucial in interaction videos: the object shape is fixed, but the hand causes rapid appearance changes, occlusions, and partial views that can otherwise corrupt per-frame reconstruction.

Freezing shape during flow inference SAM3D's geometry inference can be interpreted as predicting a velocity field for each modality (shape and layout) under conditional flow matching [6]. We initialize each frame with shared shape tokens $Z_t^S \leftarrow Z^{S\star}$ and zero out the shape update $v_\theta^S(\cdot) \equiv 0$ where v_θ^S is the learned velocity field for shape, effectively freezing the canonical shape while allowing layout to evolve. This stabilizes pose estimation even under heavy occlusion, as layout tokens can still attend to the fixed shape tokens during inference.

Temporal guidance on pose latents Even with a frozen shape, per-frame layout estimation can jitter or even flip orientation under weak visual cues such as symmetric objects and heavy occlusion. We therefore bias the per-frame layout velocity toward temporal consistency using a simple temporal guidance term. Let X_t^P be the current pose/layout latents projected from Z_t^P, and X_{t-1}^P the previous frame's latents. We modify the predicted velocity by

$$\tilde{v}_\theta^P \;=\; v_\theta^P \;-\; \lambda_{\text{temp}}\,(X_t^P - X_{t-1}^P), \tag{2}$$

which is applied to translation and rotation latents. Object scale is initialized from MV-SAM3D, refined by the first keyframe's object pointmap, and fixed during per-frame pose estimation with SAM3D to avoid frame-to-frame scale drift. Intuitively, Equation (2) discourages abrupt latent changes unless strongly supported by the current observation. To handle occasional orientation flips in the output after guidance, we perform a simple quaternion consistency check and reverse the out-of-order flips. At the end of Stage 2, we obtain an initial object trajectory $\{R_t^o, t_t^o, s^o\}_{t=1}^T$.

4.3 Stage 3: Hand–Object Tracking with SoG

So far, we have reconstructed the object shape with MV-SAM3D and arrived at an initial estimate of the object trajectory using our proposed Temporal

Guidance on pose latents. Our formulations have been agnostic of the presence of the hand mesh. Additionally, the estimated object geometry is not guaranteed to be positioned appropriately in the scene. Therefore, we now jointly refine the object and hand trajectories in short windows $\mathcal{W}$ by solving a compact tracking objective. At a high level, we maximize an image alignment energy E between a set of object Gaussians and the current frame approximated to Gaussians, while regularizing with lightweight geometric and interaction priors:

$$\begin{aligned}\min_{\mathcal{X}} \; & -\sum_{t\in\mathcal{W}} E + \lambda_{\text{j2d}}\mathcal{L}_{\text{j2d}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} \\ & + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + c^{ho}\lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}}\end{aligned} \tag{3}$$

where $\mathcal{L}_{\text{j2d}}$ is the hand-joint loss, $\mathcal{L}_{\text{depth}}$ is the depth alignment loss, $\mathcal{L}_{\text{sil}}$ is the silhouette loss, $\mathcal{L}_{\text{contact}}$ is the contact loss, and $\mathcal{L}_{\text{smooth}}$ is the smoothness loss. $\mathcal{X} = \{R_t^o, \tau_t^o, s^o, \boldsymbol{\Theta}_t, R_t^h, \tau_t^h, s^h\}_{t\in\mathcal{W}}$ denotes the variables optimized in window $\mathcal{W}$. Critical to our approach is the image-object alignment term E, which is computed using Sum-of-Gaussians (SoG) tracking. By representing the object with a compact set of 3D Gaussians and encoding the masked object image with a compact set of 2D Gaussians, SoG tracking seeks to maximize their continuous overlap similarity. We next define each ingredient of E.

For each frame t, we (i) build an *image SoG* from the RGB frame inside the object mask, and (ii) project a compact object Gaussians, *object SoG*, into the image under the current object pose.

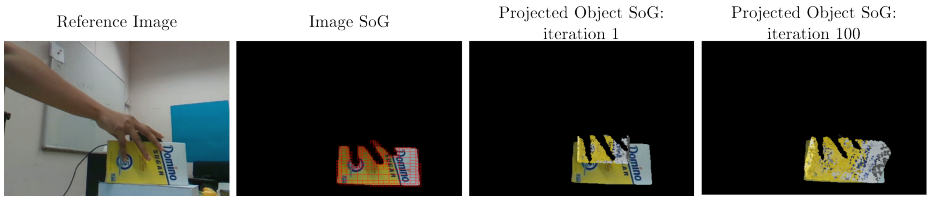

Fig. 3: Approximating the image as Gaussians and projected object SoG. From left to right: Reference frame, quad-tree build over the object area where each square represents a Gaussian, projected object SoG before optimization, projected object SoG at the end of optimization.

Image SoG Given an image I_t, we seek a compact SoG $\mathcal{K}_{I_t}$ that represents coherent pixel regions inside the object mask M_t^o. A naïve construction would assign one Gaussian to each pixel, but this is prohibitively expensive. Instead, we build a quad-tree over the object mask area to cluster pixels with similar color as shown in Fig. 3 (see Stoll et al. [49] for details).

Object SoG We decode the canonical shape tokens $Z^{S\star}$ into a dense 3D Gaussian asset via SAM3D's Gaussian decoder. Generally, these photorealistic assets are very dense with hundreds of thousands of Gaussians, which is unnecessary for tracking. We therefore sparsify the decoded asset using farthest-point sampling and keep 2000 Gaussians. The resulting object representation is a set of

3D Gaussians $\mathcal{G}_o = \{(\mu_j, \Sigma_j, c_j)\}_{j=1}^{2000}$ where $\mu_j \in \mathbb{R}^3$ is the mean, Σ_j is the diagonal covariance, and c_j is the color. For faster convergence, we use an isotropic approximation by replacing each diagonal covariance with its average variance.

SoG alignment term Now, we define the alignment energy as

$$E(\mathcal{K}_{I_t}, \Psi_t(\mathcal{G}_o)) = \sum_{i \in \mathcal{K}_{I_t}} \min\Big(\sum_{j \in \Psi_t(\mathcal{G}_o)} \upsilon_{j,t}\, E_{ij},\ E_{ii} \Big). \tag{4}$$

where $\Psi_t(\mathcal{G}_o)$ denotes the 2D object Gaussians obtained by projecting $\mathcal{G}_o$ into frame t using the current object pose and camera. The term E_{ij} is the color-weighted overlap between an image Gaussian i and a projected model Gaussian j, and E_{ii} is the self-overlap of i used for occlusion handling. A key difference from Gaussian Splatting is the simplified occlusion handling, which allows computing Gaussian similarities without sorting and in parallel (sums) instead of sequential, back-to-front alpha blending. Finally, $\upsilon_{j,t} \in \{0, 1\}$ gates out object Gaussians that are occluded by the hand (see subsection *Hand-occlusion gating*).

Pairwise overlap for two 2D Gaussians $\mathcal{B}_i$ and $\mathcal{B}_j$;

$$E_{ij} = d(c_i, c_j) \int_{\Omega} \mathcal{B}_i(u)\, \mathcal{B}_j(u)\, du, \tag{5}$$

where (c_i, c_j) are the colors, $\Omega \subset \mathbb{R}^2$ is the image plane and $d(c_i, c_j)$ is a color kernel (see the Supplementary Material for details). With isotropic covariances, the overlap has closed form:

$$\int_{\Omega} \mathcal{B}_i(u)\mathcal{B}_j(u)\, du = 2\pi \frac{\sigma_i^2 \sigma_j^2}{\sigma_i^2 + \sigma_j^2} \exp\Big(- \frac{\|\mu_i - \mu_j\|^2}{\sigma_i^2 + \sigma_j^2} \Big). \tag{6}$$

Hand-occlusion gating Different from the full-body tracking setting of Stoll et al. [49], for hand-object interaction, hand masks provide additional cues for object occlusion. Hence, rather than forcing occluded-object Gaussians to match image statistics, we restrict object-to-image matching to those visible under the hand mask. Let $\mu_{j,t}^{2D}$ be the projected mean of object Gaussian j in frame t and M_t^h be the binary hand mask. We set $\upsilon_{j,t} = \mathbf{1}\big[M_t^h\big(\mathrm{round}(\mu_{j,t}^{2D})\big) = 0\big]$, so that occluded object Gaussians do not contribute to the alignment. This prevents the optimizer from drifting the object toward unrelated visible regions when partially hidden by the hand.

Geometric and interaction priors Dyn-HaMR [61] already provides strong temporal hand tracking and predicts per-frame hand poses $\{\boldsymbol{\Theta}_t, R_t^h, \tau_t^h\}_{t=1}^N$, where $\boldsymbol{\Theta}_t$ denotes articulated joint pose parameters, R_t^h the global root orientation, and τ_t^h the global translation. Therefore, we do not employ SoG tracking for the hand; instead, we regularize the initialized MANO trajectory with lightweight geometric priors: 2D joint reprojection and a depth prior from the

initialized point map P. Using the MANO joint regressor J, we can obtain the 3D location of joint ℓ via $J_{t,\ell}(\boldsymbol{\Theta}_t, R_t^h, \tau_t^h, s^h) \in \mathbb{R}^3$. Given detected 2D joints $\hat{u}_{t,\ell}$, we use $\mathcal{L}_{\text{j2d}} = \sum_{t\in\mathcal{W}} \sum_{\ell} \left\| \Pi\big(J_{t,\ell}(\boldsymbol{\Theta}_t, \mathbf{R}_t^h, \mathbf{t}_t^h)\big) - \hat{u}_{t,\ell} \right\|_2^2$.

We further stabilize depth under occlusion by aligning rendered and pointmap depths inside the object and hand masks: D_t^{rend} for the rendered depth and D_t^{pm} for depth from the point map, a robust depth alignment term is:

$$\begin{aligned}\mathcal{L}_{\text{depth}} = \sum_{t\in\mathcal{W}} \Big(&\Big|\text{mean}\big(D_t^{\text{rend}}[M_t^o]\big) - \text{median}\big(D_t^{\text{pm}}[M_t^o]\big)\Big| \\ &+ \Big|\text{mean}\big(D_t^{\text{rend}}[M_t^h]\big) - \text{median}\big(D_t^{\text{pm}}[M_t^h]\big)\Big|\Big).\end{aligned} \tag{7}$$

This stabilizes scale/translation drift and helps under occlusions.

Additionally, to prevent degenerate SoG solutions that inflate the projected object area so that overlap energy increases, we add a silhouette loss $\mathcal{L}_{\text{sil}}$ between the rendered object silhouette and the object mask M_t^o. We also add a contact loss $\mathcal{L}_{\text{contact}}$ that penalizes the nearest-neighbor distances from MANO contact-zone vertices to the object mesh (decoded from $Z^{S\star}$), encouraging plausible hand–object proximity; this term is applied only on frames where the per-frame contact flag $c^{ho} = 1$. Lastly, to penalize temporal acceleration in translation and rotation, and suppress jitter, we add $\mathcal{L}_{\text{smooth}}$.

This optimization is lightweight because both the image and the object are compact SoGs: the image is summarized by a quad-tree of a few thousand 2D Gaussians, and the object is a ≤ 2000-Gaussian model after farthest point sampling. This revives the efficiency of classical SoG tracking [49], while leveraging generative 3D models to obtain an accurate canonical object and a strong initialization in the wild.

5 Experiments

5.1 Datasets

HO3Dv3 Dataset: HO3Dv3 [17] contains monocular videos of single-hand object interactions. In our experiments, we use the 18 sequences provided in the HOLD [14] evaluation set.

HOT3D: HOT3D [2] is an egocentric dataset with accurate 3D poses and shapes of hands and objects. Since the dataset does not provide ground-truth annotations for its test set, we instead select 18 sequences from the training set. Specifically, we extract segments from the full training videos that contain only single hand–object interactions with different objects. Before running the methods on HOT3D, the fisheye videos are undistorted.

5.2 Metrics

We decode the shape tokens into object meshes using SAM3D's mesh decoder, and then follow the evaluation protocol in HOLD [14]. We use root-relative mean-per-joint position error (MPJPE, in mm) to measure hand pose accuracy, and

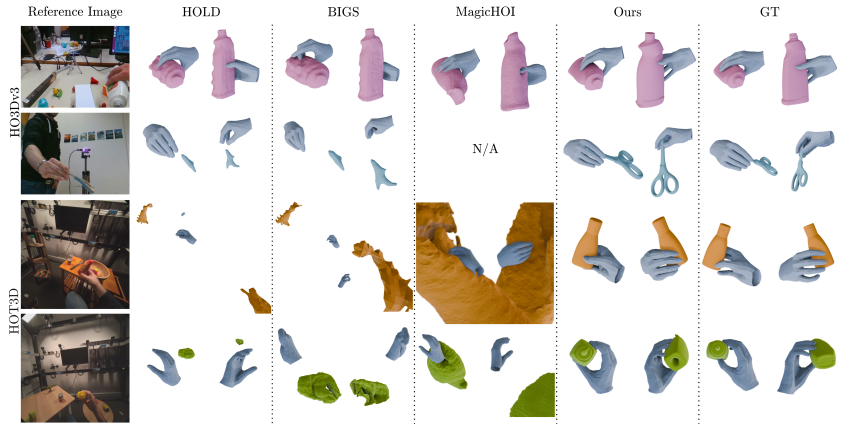

Fig. 4: Qualitative Comparison. We compare the output of GraG with previous SoTA works HOLD, BIGS, and MagicHOI on HO3D (first 2 rows) and HOT3D (last 2 rows). In the 2nd row 4th column, MagicHOI fails to produce a valid reconstruction; we therefore report it as `N/A`. Overall, GraG preserves sharper object geometry and yields more plausible hand poses (with fewer interpenetrations), while being substantially more efficient. For more qualitative results, please refer to the Supplementary Material.

Chamfer distance (CD, in cm) and F-score (in %) to evaluate object reconstruction quality. To evaluate object template quality independent of object pose, following HOLD [14], we ICP-align the predicted object mesh to the ground-truth mesh and compute CD and F-score at 10 mm (F10). To measure object pose and shape relative to the hand in 3D, we subtract the predicted hand root from the object vertices and compute the hand-relative CD for the object (CD_h) [14]. We also report Success Rate (SR) where a sequence is deemed as a failure if the method fails to output a valid result such that when initialization breaks down (notably for SfM-based pipelines) or when tracking degrades severely such that the hand-relative CD reaches 1000 cm. Since methods are run with different clip lengths, we report normalized runtime (hours per 100 frames) on a single NVIDIA RTX4090 GPU to quantify efficiency, including preprocessing and optimization.

5.3 Implementation Details

We optimize each sequence with AdamW [32] in a sliding-window manner: running 100 iterations per window of 8 frames (stride 1), sequentially covering the full video. Within each window, we optimize the hand and object pose parameters, while keeping the canonical object and hand shape fixed after initialization. On a single RTX4090 GPU, for a 100-frame sequence, Stages 1-2 take ∼6 minutes on average and Stage 3 takes ∼30 minutes. We report all hyperparameters in the Supplementary Material.

Table 1: Quantitative comparison. We report object reconstruction, hand accuracy, success rate, and runtime on HO3Dv3 and HOT3D. We highlight the best and second-best results. All methods are averaged over the *successful results* only.

Dataset	Method	CD [cm] ↓	F10 [%] ↑	MPJPE [mm] ↓	CD_h [cm] ↓	Success Rate [%] ↑	Runtime [h] ↓
HO3Dv3	HOLD [14]	0.78	92.0	23.4	**4.27**	**100**	10.5
	BIGS [34]	0.67	94.1	23.9	7.61	**100**	3.60
	MagicHOI [54]	1.58	73.7	**4.35**	129	67.0	1.20
	Ours	**0.58**	**96.7**	8.18	5.24	**100**	**0.56**
HOT3D	HOLD [14]	3.14	53.3	21.9	122	50.0	10.5
	BIGS [34]	2.92	51.7	36.9	84.2	50.0	3.60
	MagicHOI [54]	22.6	33.6	**18.4**	97.1	55.6	1.20
	Ours	**0.76**	**93.5**	21.7	**9.71**	**100**	**0.56**

5.4 Results

Compared SoTA Methods: We compare GraG to state-of-the-art methods including HOLD [14], BIGS [34], and MagicHOI [54], following HOLD's evaluation protocol. HOLD is a NeRF-based method that jointly optimizes hand–object scenes using SDF representations. BIGS reconstructs HOI scenes using 3D Gaussians with a triplane MLP and SDS loss. MagicHOI extends HOLD by leveraging novel-view synthesis models to hallucinate occluded object regions.

Quantitative Results: We report quantitative results in Tab. 1. On HO3Dv3, our method reconstructs object shapes with significantly lower error, improving CD by 13.4% over BIGS and 25.6% over HOLD, while achieving an F10 score of 96.7%. Hand pose accuracy is also significantly improved: our MPJPE is 65.0% lower than HOLD and 65.8% lower than BIGS, while MagicHOI achieves the best MPJPE in this setting. Notably, despite superior accuracy in hand and object reconstruction, our method requires only a fraction of the compute, consuming just 0.56 hours per 100 frames, compared to 1.2 for MagicHOI, 3.6 for BIGS, and 10.5 hours for HOLD.

On the more challenging HOT3D dataset, characterized by egocentric viewpoints, rapid camera motion, and heavy hand occlusions, our method further improves both object reconstruction and hand–object alignment. We reduce CD by 74.0% over BIGS and 75.8% over HOLD, and achieve the best hand-object alignment (lowest CD_h). Our method also achieves a 100% success rate, whereas SfM-based baselines drop to 50–56%. In long egocentric sequences, COLMAP-based initialization is sensitive to frame sampling and often produces unstable camera poses, which leads to drift or failed reconstructions. Consequently, methods that depend on such initialization (HOLD, BIGS, and MagicHOI) frequently fail to converge or produce inconsistent object tracking. Specifically, MagicHOI is designed for short-clip evaluation; when applied to our full-length HO3Dv3 and egocentric HOT3D sequences, its COLMAP-based initialization frequently becomes unstable, reducing its success rate.

Qualitative Results: As shown in Fig. 4, our method accurately reconstructs both objects and hands from the input video. Compared to prior SoTA methods, our reconstructions preserve finer geometric details and successfully recover objects with thin structures, such as the scissors (second row), which remain chal-

Table 2: Ablation study. We ablate key components of our method on 4 sequences. $\mathcal{L}_{\text{depth}}$ and $\mathcal{L}_{\text{contact}}$ most strongly improve hand-object relative poses (lower CD_h), while SoG refinement provides robust gains without sacrificing runtime.

Ablation	CD [cm] ↓	F10 [%] ↑	MPJPE [mm] ↓	CD_h [cm] ↓	Runtime [h] ↓
w/o keyframe selection (random K frames)	0.79	90.5	9.26	4.97	0.56
w/ single view from first frame ($K = 1$)	0.68	91.6	9.26	4.89	0.56
w/o freezing shape	0.41	99.1	9.25	5.35	0.56
w/o temporal guidance (λ_{temp}=0)	0.41	99.1	9.27	5.00	0.56
w/o Stage 3	0.41	99.1	9.19	10.9	0.10
Gaussian Splatting tracking instead of SoG	0.41	99.1	9.14	13.4	1.20
w/o SoG refinement	0.41	99.1	9.29	10.1	0.55
w/o visibility gate (no hand-occlusion gating)	0.41	99.1	9.27	5.00	0.56
w/o depth loss $\mathcal{L}_{\text{depth}}$	0.41	99.1	9.17	7.12	0.56
w/o silhouette loss $\mathcal{L}_{\text{sil}}$	0.41	99.1	9.27	4.96	0.56
w/o contact loss $\mathcal{L}_{\text{contact}}$	0.41	99.1	9.30	6.30	0.55
Full model	0.41	99.1	8.92	4.08	0.56

lenging for existing baselines. Furthermore, our approach produces substantially more stable hand–object interaction reconstructions on the HOT3D dataset. As illustrated in the third and fourth rows, competing methods either fail to reconstruct meaningful object geometry or diverge during optimization, largely due to their reliance on COLMAP/SfM initialization. In contrast, our approach remains robust under egocentric conditions where objects appear small, move rapidly, and are heavily occluded by the hand. Our method also recovers more accurate hand poses; as illustrated in the fifth column, we observe fewer instances of hand-object interpenetration and more physically plausible grasping configurations. We additionally provide qualitative visualizations by overlaying our reconstructions on the reference images in the Supplementary Material.

6 Ablations

We validate our design choices by ablating on four HO3Dv3 sequences. Results are presented in Tab. 2, Fig. 5 and the Supplementary Material. Random keyframe selection degrades object reconstruction (CD 0.79, F10 90.5), while single-view initialization (K=1) improves but still underperforms (CD 0.68, F10 91.6), supporting the need for informative multi-view initialization. In Stage 2, removing shape freezing or temporal guidance increases hand-relative object pose error, indicating reduced tracking stability. Skipping Stage 3, i.e., evaluating direct outputs of our modified SAM3D, yields high object pose error (CD_h 10.9). Removing SoG refinement similarly hurts performance (CD_h 10.1). We also highlight the efficiency of GraG by replacing our lightweight optimization with dense Gaussian tracking. In this setting, we optimize the hand and object poses using dense Gaussians with RGB loss together with our geometric and interaction priors. It yields worse object pose tracking (CD_h 13.4) and is markedly slower (1.20h vs. 0.56h), motivating our compact formulation. Among losses, depth and contact are most important while silhouette and visibility gating provide improvements under occlusion. Overall, the full model achieves the best accuracy (MPJPE 8.92, CD_h 4.08).

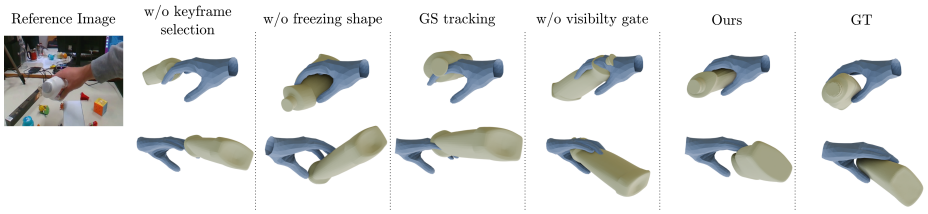

Fig. 5: Ablation experiments. We visualize how key design choices affect reconstruction quality (two representative views per setting: camera view and back view). Random keyframe selection can yield an inaccurate canonical object (shape/scale), leading to implausible grasps. Without freezing the canonical shape in our video-adapted SAM3D, per-frame pose estimates become unstable. Replacing our compact SoG refinement with dense Gaussian Splatting (GS) tracking is slower and often fails to converge under the same iteration budget. Without the visibility gate, SoG alignment is biased by hand-occluded regions, producing incorrect object poses. Our full model most closely matches the ground truth.

7 Limitations

Our method successfully integrates several foundational models with an efficient tracking approach to perform HOI reconstruction. While it is robust to minor errors, it cannot recover when the base foundational models fail completely. In particular, we find the monocular pointmap depths produced by DepthAnything 3 to be a critical component, a failure of which leads to tracking drift and incorrect depth alignment, which is demonstrated in the Supplementary Material. We also depend on good hand and object masks, which can, at times, be challenging to recover under extreme occlusion. Additionally, the current method targets a single manipulated rigid object with a single dominant interacting hand; handling multiple objects, strong hand-to-hand occlusions, or deformable objects is future work.

8 Conclusion

We introduced Grasp in Gaussians (**GraG**), an efficient and scalable framework for reconstructing dynamic 3D hand-object interactions from a single monocular video. We showed that naïvely combining the 2D/3D foundational models can produce suboptimal results, particularly in terms of optimization runtime. We addressed this by re-introducing the idea of Sum-of-Gaussian tracking and showed how a sparse set of Gaussians can be tracked efficiently, and without losing the reconstruction quality. Experiments demonstrate state-of-the-art accuracy on standard benchmarks and strong generalization to in-the-wild videos, while substantially improving efficiency compared to prior video-based HOI reconstruction methods. In future work, we aim to integrate physical plausibility of the reconstruction to improve robustness and extend the pipeline to multiple hands and interacting objects.

References

1. Aytekin, A.I., Rhodin, H., Dabral, R., Theobalt, C.: Follow my hold: Hand-object interaction reconstruction through geometric guidance. In: Thirteenth International Conference on 3D Vision (2026)
2. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7061–7071 (2025)
3. Bretzner, L., Laptev, I., Lindeberg, T.: Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. In: Automatic Face and Gesture Recognition. pp. 423–428 (2002). https://doi.org/10.1109/AFGR.2002.1004190
4. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12417–12426 (2021)
5. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
6. Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., et al.: Sam 3d: 3dfy anything in images. arXiv preprint arXiv:2511.16624 (2025)
7. Collins, J., Goel, S., Deng, K., Luthra, A., Xu, L., Gundogdu, E., Zhang, X., Vicente, T.F.Y., Dideriksen, T., Arora, H., et al.: Abo: Dataset and benchmarks for real-world 3d object understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21126–21136 (2022)
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Cong, X., Xing, A., Pokhariya, C., Fu, R., Sridhar, S.: Dytact: Capturing dynamic contacts in hand-object manipulation. arXiv preprint arXiv:2506.03103 (2025)
10. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5031–5041 (2020)
11. Deitke, M., Liu, R., Wallingford, M., Ngo, H., Michel, O., Kusupati, A., Fan, A., Laforte, C., Voleti, V., Gadre, S.Y., et al.: Objaverse-xl: A universe of 10m+ 3d objects. Advances in Neural Information Processing Systems **36**, 35799–35813 (2023)
12. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
13. Elhayek, A., Aguiar, E., Jain, A., Tompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Efficient ConvNet-based marker-less motion capture in general scenes with a low number of cameras. In: CVPR (2015)
14. 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)

15. 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)
16. Hampali, S., Hodan, T., Tran, L., Ma, L., Keskin, C., Lepetit, V.: In-hand 3d object scanning from an rgb sequence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17079–17088 (2023)
17. Hampali, S., Sarkar, S.D., Lepetit, V.: Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset. arXiv preprint arXiv:2107.00887 (2021)
18. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge, England, 2 edn. (Jan 2011)
19. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 571–580 (2020)
20. Hasson, Y., Varol, G., Schmid, C., Laptev, I.: Towards unconstrained joint hand-object reconstruction from rgb videos. In: 2021 International Conference on 3D Vision (3DV). pp. 659–668. IEEE (2021)
21. 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)
22. Huang, D., Ji, X., He, X., Sun, J., He, T., Shuai, Q., Ouyang, W., Zhou, X.: Reconstructing hand-held objects from monocular video. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
23. Karunratanakul, K., Yang, J., Zhang, Y., Black, M.J., Muandet, K., Tang, S.: Grasping field: Learning implicit representations for human grasps. In: 2020 International Conference on 3D Vision (3DV). pp. 333–344. IEEE (2020)
24. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
25. Lai, Z., Zhao, Y., Liu, H., Zhao, Z., Lin, Q., Shi, H., Yang, X., Yang, M., Yang, S., Feng, Y., et al.: Hunyuan3d 2.5: Towards high-fidelity 3d assets generation with ultimate details. arXiv preprint arXiv:2506.16504 (2025)
26. Li, B.: MV-SAM3D: SAM 3d objects with multi-view images. GitHub repository (January 2025), `https://github.com/devinli123/MV-SAM3D`
27. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
28. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9298–9309 (2023)
29. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3d hand-object poses estimation with interactions in time. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14687–14697 (2021)
30. 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)

31. Liu, Y., Long, X., Yang, Z., Liu, Y., Habermann, M., Theobalt, C., Ma, Y., Wang, W.: Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7037–7047 (2025)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
33. Moeslund, T.B., Hilton, A., Krüger, V.: A survey of advances in vision-based human motion capture and analysis. CVIU **104**(2), 90–126 (2006). https://doi.org/10.1016/j.cviu.2006.08.002
34. 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. pp. 17437–17447 (2025)
35. Plankers, R., Fua, P.: Articulated soft objects for multiview shape and motion capture. PAMI **25**(9), 1182–1187 (2003)
36. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022)
37. Rajič, F., Xu, H., Mihajlovic, M., Li, S., Demir, I., Gündoğdu, E., Ke, L., Prokudin, S., Pollefeys, M., Tang, S.: Multi-view 3d point tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 59–68 (2025)
38. Ren, C.Y., Prisacariu, V., Kaehler, O., Reid, I., Murray, D.: 3D tracking of multiple objects with identical appearance using RGB-D input. In: 3DV. pp. 47–54 (2014). https://doi.org/10.1109/3DV.2014.39, http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7035808
39. Rhodin, H., Richardt, C., Casas, D., Insafutdinov, E., Shafiei, M., Seidel, H.P., Schiele, B., Theobalt, C.: Egocap: egocentric marker-less motion capture with two fisheye cameras. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016)
40. Rhodin, H., Robertini, N., Casas, D., Richardt, C., Seidel, H.P., Theobalt, C.: General automatic human shape and motion capture using volumetric contour cues. In: European conference on computer vision. pp. 509–526. Springer (2016)
41. Rhodin, H., Robertini, N., Richardt, C., Seidel, H.P., Theobalt, C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In: Proceedings of the IEEE international conference on computer vision. pp. 765–773 (2015)
42. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610 (2022)
43. Sarlin, P.E., Cadena, C., Siegwart, R., Dymczyk, M.: From coarse to fine: Robust hierarchical localization at large scale. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12716–12725 (2019)
44. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4938–4947 (2020)
45. Scharstein, D., Szeliski, R., Zabih, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. In: Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001). IEEE Comput. Soc (2002)
46. Sridhar, S., Mueller, F., Oulasvirta, A., Theobalt, C.: Fast and robust hand tracking using detection-guided optimization. In: CVPR (2015), http://handtracker.mpi-inf.mpg.de/projects/FastHandTracker/
47. Sridhar, S., Rhodin, H., Seidel, H.P., Oulasvirta, A., Theobalt, C.: Real-time hand tracking using a sum of anisotropic gaussians model. In: 2014 2nd International Conference on 3D Vision. vol. 1, pp. 319–326. IEEE (2014)

48. Stoiber, M., Pfanne, M., Strobl, K.H., Triebel, R., Albu-Schäffer, A.: Srt3d: A sparse region-based 3d object tracking approach for the real world. International Journal of Computer Vision **130**(4), 1008–1030 (2022)
49. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: 2011 international conference on computer vision. pp. 951–958. IEEE (2011)
50. Szeliski, R.: Computer Vision. Texts in computer science, Springer, London, England (Oct 2010)
51. Team, V., Hong, W., Yu, W., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
52. Tekin, B., Bogo, F., Pollefeys, M.: H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4511–4520 (2019)
53. Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. Int. J. Comput. Vis. **9**(2), 137–154 (Nov 1992)
54. 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. arXiv preprint arXiv:2508.05506 (2025)
55. Wu, G., Fang, J., Yang, C., Li, S., Yi, T., Lu, J., Zhou, Z., Cen, J., Xie, L., Zhang, X., Wei, W., Liu, W., Wang, X., Tian, Q.: Unilat3d: Geometry-appearance unified latents for single-stage 3d generation (2025), `https://arxiv.org/abs/2509.25079`
56. Xiang, J., Lv, Z., Xu, S., Deng, Y., Wang, R., Zhang, B., Chen, D., Tong, X., Yang, J.: Structured 3d latents for scalable and versatile 3d generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21469–21480 (June 2025)
57. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: Cpf: Learning a contact potential field to model the hand-object interaction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11097–11106 (2021)
58. Ye, C., Wu, Y., Lu, Z., Chang, J., Guo, X., Zhou, J., Zhao, H., Han, X.: Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. arXiv preprint arXiv:2503.22236 **3**, 2 (2025)
59. Ye, Y., Gupta, A., Tulsiani, S.: What's in your hands? 3d reconstruction of generic objects in hands. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3895–3905 (2022)
60. 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. pp. 19717–19728 (2023)
61. Yu, Z., Zafeiriou, S., Birdal, T.: Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 27716–27726 (2025)
62. Zheng, C., Xue, L., Zarate, J., Song, J.: Gaustar: Gaussian surface tracking and reconstruction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16543–16553 (2025)
63. Zhong, L., Yang, L., Li, K., Zhen, H., Han, M., Lu, C.: Color-neus: Reconstructing neural implicit surfaces with color. In: 2024 International Conference on 3D Vision (3DV). pp. 631–640. IEEE (2024)
64. Zhong, Y., Jain, A.K., Dubuisson-Jolly, M.P.: Object tracking using deformable templates. IEEE transactions on pattern analysis and machine intelligence **22**(5), 544–549 (2000)