RECONSTRUCTING DEFORMABLE HAND-OBJECT MANIPULATION VIA VISION-PHYSICS INVERSE MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing dynamic hand-object interaction from visual observations is fundamental for enabling embodied agents to acquire dexterous manipulation skills. While existing methods have achieved promising results under rigid-body assumptions or template-based representations, they struggle to capture the complex, spatially varying deformations arising in interactions with deformable objects, especially in bimanual settings. In this paper, we present a novel vision-physics inverse modeling framework for reconstructing bimanual dynamic deformable hand-object interactions from multi-view videos. Our key insight is that object deformation is a physical response to hand-applied forces, which is directly reflected in temporal appearance variations across views. Building on this observation, our method jointly leverages geometry, appearance, and interaction cues from multi-view observations to support physics-constrained non-rigid reconstruction. Specifically, we introduce a physics-aware deformation model that estimates both a time-varying force field and a time-invariant material field, allowing fine-grained deformations to be recovered via differentiable simulation. To facilitate evaluation, we further construct a real-world multi-view dataset of bimanual interactions with diverse deformable objects. Extensive experiments demonstrate that our method significantly outperforms prior approaches in both geometric accuracy and interaction consistency, particularly in challenging deformable interaction scenarios. The code and dataset will be made publicly available.

1 INTRODUCTION

Reconstructing 3D hand-object interaction (HOI) from visual observations serves as a cornerstone for building manipulation skills in embodied intelligence (Cao et al., 2021; Huang et al., 2022; Swamy et al., 2023; Yang et al., 2023b; Wang et al., 2023; Zhang et al., 2024b; Fan et al., 2024a; Pokhariya et al., 2024; Tian et al., 2025; Jiang et al., 2025b). To acquire human-level dexterous manipulation skills, embodied agents must not only mimic kinematic hand trajectories, but also deeply understand the dynamic physical responses of the objects being manipulated.

Recent works (On et al., 2025; Zhang et al., 2024a; Wang et al., 2025) have recovered visually plausible geometries and motions for hand-object interactions by assuming object rigidity (Ye et al., 2022; Fan et al., 2024b; Chen et al., 2022; Wu et al., 2024; Qu et al., 2023; Pang et al., 2024; Ye et al., 2024; Prakash et al., 2024; Baek et al., 2020; Fu et al., 2024; Cha et al., 2024; Cho et al., 2023; Chen et al., 2025) or adopting category-level templates (Baek et al., 2020; Zhu et al., 2024; Liu et al., 2022; Yang et al., 2022; Zuo et al., 2023; Li et al., 2023). Yet, these simplified representations cannot capture the evolving physical dynamics of deformable objects during hand-object interaction. A few recent deformable HOI methods (Xie et al., 2023; Yu et al., 2025; Hu et al., 2022) relax the rigidity assumption through non-rigid optimization, but typically rely on coarse discretizations or low-dimensional deformation models. This limits their capacity to model high-dimensional, spatially varying deformations driven by complex bimanual contacts.

During bimanual manipulation of deformable objects, object deformation arises as a physical response to hand-applied contact forces, while the resulting dynamics are directly reflected in contin-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

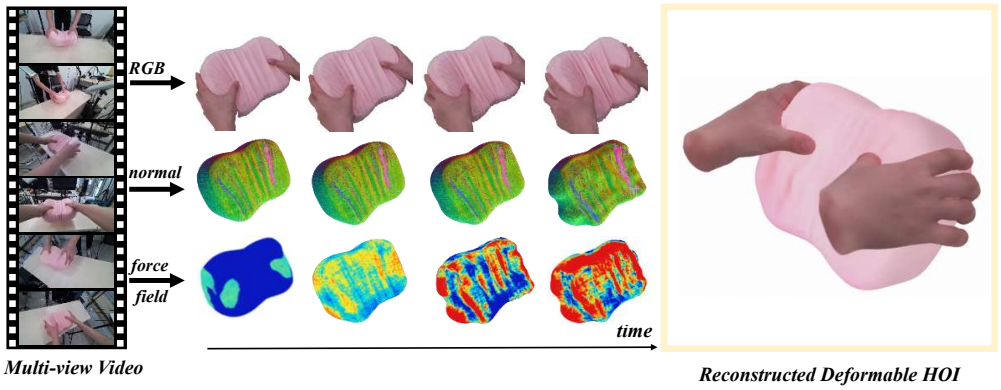


Figure 1: During bimanual manipulation of deformable objects, hand-applied contact forces drive non-rigid object deformation, while the resulting physical dynamics are reflected in the object's evolving visual appearance across multi-view observations.

uous changes in the object's visual appearance, as shown in Fig. 1. This coupling between physical deformation and visual appearance motivates our vision-physics inverse modeling framework, which jointly estimates a time-varying force field and a time-invariant material field to recover non-rigid bimanual hand-object interactions from multi-view observations.

Specifically, our method consists of two stages. First, we use foundation models (Li et al., 2026; Ravi et al., 2024; Potamias et al., 2025) to initialize coarse hand-object geometry and object appearance from multi-view observations. We further adopt a 3D Gaussian Splatting (3DGS) representation (Tian et al., 2025; On et al., 2025; Kerbl et al., 2023; Xie et al., 2024) for these initial estimates. By leveraging multi-view appearance cues through differentiable rendering, this stage refines the rigid hand and object poses and provides geometry, appearance, and interaction priors for subsequent physics-aware non-rigid reconstruction. Second, we recover fine-grained non-rigid object deformation through physics-based inverse modeling of multi-view appearance changes. Instead of directly warping geometry with heuristic deformation models, we estimate the underlying time-varying force field and time-invariant material field that explain the observed visual variations. These physical fields drive a deformation model that reconstructs fine-grained geometry under complex, changing contact states.

Furthermore, existing HOI datasets are predominantly biased toward rigid objects or simplified single-hand non-rigid scenarios (Hampali et al., 2019; Chao et al., 2021; Taheri et al., 2020; Banerjee et al., 2025), falling short of supporting training and evaluation for complex bimanual deformable HOI. Our reconstruction method enables scalable annotation of such interactions from real-world multi-view videos, allowing us to construct and release a dataset of bimanual interactions with diverse elastic objects and establish a novel benchmark for dynamic deformable HOI reconstruction under dual-hand manipulation. In conclusion, our contributions are threefold:

- We present a novel framework for bimanual deformable hand-object interaction reconstruction from multi-view videos, where hand-induced non-rigid deformation is inferred from evolving object appearance to jointly recover time-varying geometry and dual-hand interaction states.
- We propose a non-rigid vision-physics inverse modeling framework that uses multi-view appearance changes as the driving signal. It jointly estimates a time-varying force field and a time-invariant material field, and integrates them into differentiable non-rigid simulation to reconstruct dynamic deformations consistent with observed visual evidence without relying on category templates.
- Enabled by our reconstruction framework, we construct and release a real-world multi-view dataset of bimanual interactions with diverse deformable objects, providing a challenging evaluation resource for research on dynamic deformable HOI reconstruction.

2 RELATED WORK

Hand-object interaction reconstruction. Early HOI reconstruction methods primarily focus on single-hand manipulation. Representative works such as Ye et al. (2022) enable category-agnostic object reconstruction from monocular images, while follow-ups (Chen et al., 2022; 2023; Qi et al., 2024; Zhang et al., 2023) improve hand-object alignment via SDF-based constraints. More recent approaches improve robustness in unconstrained settings through diffusion priors (Ye et al., 2023; Yang et al., 2023a), foundation models (Liu et al., 2025), and occlusion-aware frameworks (Zhang et al., 2024a; Wu et al., 2024; Swamy et al., 2025). However, these methods largely assume rigid object geometry. To address more complex interactions, subsequent works extend HOI reconstruction to bimanual settings (Tawfik Aboukhadra et al., 2022; Ren et al., 2023), where occlusions and contact patterns become significantly more challenging. Recent advances further integrate generative priors, particularly diffusion-based models. Recent pipelines (Fan et al., 2024b; On et al., 2025; Aytekin et al., 2025) achieve category-agnostic bimanual reconstruction using Gaussian splatting and geometry-guided constraints. Despite their progress, they still rely on rigid or template-based object representations. As a result, a critical gap remains between existing HOI systems and real-world interactions involving deformable objects. While a few pioneering works explore non-rigid object reconstruction (Xie et al., 2023; Yu et al., 2025), they typically depend on specialized sensors or simplified interaction scenarios. In contrast, our method jointly reconstructs bimanual hand motions and interaction-driven deformations of non-rigid objects, bridging visual reconstruction and physics-constrained interaction modeling.

Non-rigid object modeling. Modeling non-rigid object deformation has gained increasing attention through implicit representations, differentiable physics, and Gaussian-based dynamic modeling. Implicit deformation fields (Pumarola et al., 2021; Park et al., 2021a; Zhang et al., 2024d), such as HyperNeRF (Park et al., 2021b), extend radiance fields into higher-dimensional embeddings to capture topological variations, while DynoSurf (Yao et al., 2024) learns deformation-aware dynamic surface representations from videos. In parallel, differentiable physics methods, including DiffPD (Du et al., 2021) and NeuSpring (Xu et al., 2025), introduce physics-based formulations for modeling soft-body dynamics. More recently, Gaussian-based methods (Rong et al., 2025; Liu et al., 2024a) represent deformable objects through time-varying 3D Gaussians. PhysGaussian (Xie et al., 2024) integrates MPM-based physical simulation with 3D Gaussians, and subsequent methods (Zhang et al., 2024c; Huang et al., 2024) adopt related formulations to model deformable objects. Meanwhile, PhysTwin (Jiang et al., 2025a) reconstructs deformable object dynamics using a differentiable spring-mass model. Our method instead recovers the complete hand-object manipulation state, including articulated hand poses, global hand-object alignment, contact configurations, and time-varying object deformation. This joint reconstruction is particularly challenging due to the severe occlusions introduced by hand-object interaction, but provides richer interaction representations for downstream embodied-AI applications, such as manipulation understanding and imitation learning.

Hand dataset. Existing HOI datasets have evolved in scale, annotation quality, and interaction diversity. Early datasets such as FPHA (Jiang et al., 2021), FreiHAND (Zimmermann et al., 2019), HO-3D (Hampali et al., 2019), and GanHand (Corona et al., 2020) focus primarily on single-hand interactions with rigid objects. Subsequent large-scale datasets, including ContactPose (Brahmbhatt et al., 2020), GRAB (Taheri et al., 2020), DexYCB (Chao et al., 2021), HOI4D (Liu et al., 2022), the OakInk series (Yang et al., 2022; Zhan et al., 2024), ContactArt (Zhu et al., 2024), HOGraspNet (Cho et al., 2024), TACO (Liu et al., 2024b), HOT3D (Banerjee et al., 2025), and GigaHands (Fu et al., 2025), further expand viewpoint, subject, and interaction diversity, with some providing detailed contact annotations. However, datasets specifically targeting deformable-object interactions remain scarce. Existing attempts such as HMDO (Xie et al., 2023) and HOT (Yu et al., 2025) include deformable objects but remain limited in scale, diversity, or deformation realism. Overall, current HOI datasets provide broad coverage for rigid-object scenarios, yet lack high-quality, multi-view, temporally aligned data for reconstructing realistic deformable interactions.

3 METHOD

Our method aims to jointly recover the hand-object interaction states and the non-rigid deformation of the object from a multi-view video sequence that captures the interaction between two hands and a

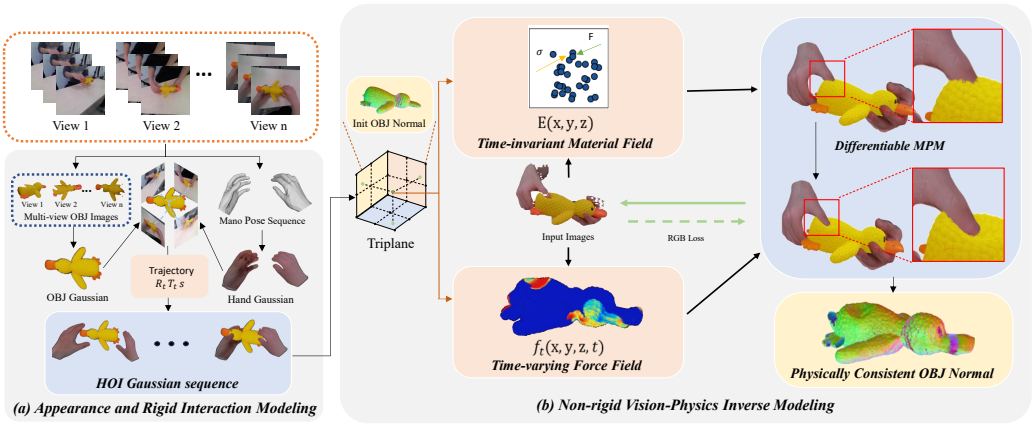


Figure 2: **Overview of our framework.** Our method decomposes reconstruction into two stages. (a) Appearance and Rigid Interaction Modeling initializes hand-object geometry from multi-view videos and represents it with 3D Gaussians, using appearance cues to refine rigid poses. (b) Non-rigid Vision-Physics Inverse Modeling parameterizes a time-invariant material field and a time-varying force field with triplanes, then optimizes them through multi-view RGB supervision and differentiable Material Point Method (MPM) simulation to recover physically consistent object deformations over time.

deformable object. During the initialization phase, given a multi-view hand-object interaction video $V = \{I_t\}_{t=1}^{T}$ and the corresponding camera parameters, we first segment the hands (Potamias et al., 2025) and the object (Ravi et al., 2024) in each view, reconstruct the object mesh using MV-SAM3D (Li et al., 2026), and separately optimize the initial hand-object motion sequences. Subsequently, we represent the hands and the object individually using 3D Gaussians (Kerbl et al., 2023), and perform non-rigid optimization of the hand-object interaction under appearance constraints to obtain the refined hand-object pose sequence. We define the time-varying force field exerted on the object (Zhang et al., 2024c) during the hand-object interaction using temporal triplane parameterization, and optimize the force field by driving Gaussian motion through the MPM process, thereby obtaining the deformed Gaussians of the object at each frame.

3.1 Hand-object initialization

Before reconstructing deformable hand–object interactions, we first initialize the interaction poses of the hand and the object to provide a reliable prior for subsequent Gaussian-based modeling and physics-aware optimization. Specifically, given synchronized multi-view RGB videos V, we first adopt LangSAM to perform instance-level segmentation of hands and objects in each view, yielding hand–object masks $\left\{M_t^h, M_t^o\right\}_{t=1}^{T}$. We then calculate the intersection over union (IoU) of the hand and object in each view, select the object segmentation image from the view with the minimum overlap, and employ Multi-View SAM 3D Object (Li et al., 2026) to obtain the object's mesh template and initial Gaussians. Using the object mesh and its corresponding mask, we further optimize the 3D translation, rotation, and scale of the object via multi-view optimization. Meanwhile, we reconstruct the initial pose sequence of the hand from multiple views using the HoCap annotation method (Wang et al., 2024).

3.2 Appearance and rigid interaction modeling

Given the initialized hand pose sequence, we represent the hand using MANO-based Gaussians and jointly optimize the hand pose and appearance parameters under multi-view supervision. During optimization, the hand Gaussian topology is kept fixed, while the hand pose sequence and Gaussian appearance are refined using photometric losses between rendered views and observations.

For the object, we initialize a Gaussian representation from the reconstructed object geometry and jointly optimize it with the hand Gaussians under a unified Gaussian representation. Given an initial rigid transformation parameterized by translation, rotation, and scale, the hand-object Gaussians are rendered and compared with the observed multi-view images using photometric supervision. The

rigid transformation is optimized to improve the alignment between the reconstructed object and the observed hand-object interaction before applying the subsequent non-rigid deformation optimization. Details of the hand Gaussian initialization are provided in the supplementary material.

3.3 NON-RIGID VISION-PHYSICS INVERSE MODELING

To model the non-rigid deformation of elastic objects, we consider deformation as the result of the interplay between external forces and internal stresses. Instead of directly optimizing geometry, we optimize external forces that drive the motion of object Gaussians, thereby inducing deformation. To ensure physical plausibility, we incorporate a differentiable MPM simulation (Xie et al., 2024; Huang et al., 2024; Zhang et al., 2024c) to constrain the optimization. The recovered force and material fields are effective latent simulator parameters inferred from visual observations, rather than direct measurements of absolute contact forces or material properties. Their validity is evaluated through observable deformation geometry and interaction consistency.

Time-varying force field. Existing methods typically assume time-invariant physical quantities (e.g., initial velocity or material fields), and rely on MPM constraints to generate dynamic deformations governed by inertia. However, in hand–object interaction scenarios, the forces applied by the hand vary over time, and the resulting deformation cannot be explained by inertia alone. To address this limitation, we propose a time-varying force field that models the spatiotemporal distribution of external forces throughout the interaction. Specifically, we parameterize the force field using a temporal triplane representation followed by a three-layer MLP. This field is designed to capture spatially and temporally varying interaction loads that drive compression- and tension-dominated deformations under sustained contact, including pressing, squeezing, pulling, stretching, compression, and bending. The inferred forces serve as effective latent driving signals that may absorb the normal and tangential effects required to explain the observed deformation.

Time-invariant material field. To simulate the stress corresponding to the time-varying force field, we optimize a time-invariant field to represent the effective material properties. The formulation of the material field is detailed in the supplementary material. We parameterize the material field using a standard triplane representation followed by a three-layer MLP, and impose spatial smoothness regularization to ensure continuity of material distribution.

Physics-aware non-rigid deformation. When the hand interacts with the object, we detect contact regions based on the spatial proximity between hand Gaussians and object Gaussians. External forces are then applied to object Gaussians within these regions to simulate hand-induced interactions.

Vision-physics constrained optimization.

We jointly optimize the hand pose sequence, the time-varying force field, and the material field under visual, physical, and interaction constraints. For each frame, the object Gaussians are propagated through the MPM simulation under the estimated external forces and internal material responses, producing updated positions and covariances. The posed hand Gaussians and deformed object Gaussians are then jointly rendered from multiple views and compared with the segmented hand–object observations. Because the contact loss is evaluated between the current hand and deformed object Gaussians, its gradients update both the hand pose and the physical fields, enabling contact-mediated bidirectional refinement of the hand and object reconstructions.

The optimization objective is formulated as:

$$\underset{\{\theta_t\}_{t=1}^{T},\,\Phi_f,\,\Phi_m}{\text{minimize}} \quad \mathcal{L} = \lambda_i \mathcal{L}_{\text{img}} + \lambda_c \mathcal{L}_{\text{contact}} + \lambda_r \mathcal{L}_{\text{reg}}, \tag{1}$$

where $\{\theta_t\}_{t=1}^{T}$ denotes the hand pose sequence, and Φ_f and Φ_m denote the learnable parameters of the time-varying force field and the time-invariant material field, respectively.

$$\mathcal{L}_{\text{img}} = |I_{\text{render}} - I_{\text{obs}}|_2^2 + \alpha\,(1 - \text{SSIM}(I_{\text{render}}, I_{\text{obs}}))\,. \tag{2}$$

This term enforces consistency between the rendered Gaussian images I_{render} and the observed interaction images I_{obs} at both pixel and structural levels.

We compute the contact loss using bidirectional nearest-neighbor distances over stochastically subsampled object and hand Gaussians, denoted as $\tilde{\mathcal{G}}_o$ and $\tilde{\mathcal{G}}_h$, respectively. For an object point $x_i \in \tilde{\mathcal{G}}_o$, its shortest distance to the hand is defined as $d_{i\to h} = \min_{x_j \in \tilde{\mathcal{G}}_h} \|x_i - x_j\|_2$. The hand-to-object distance $d_{j\to o}$ is defined symmetrically.

The overall interaction objective comprises two distinct terms: a penetration penalty $\mathcal{L}_{\text{pen}}$ and an attractive contact term $\mathcal{L}_{\text{attr}}$:

$$\mathcal{L}_{\text{contact}} = \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}}. \tag{3}$$

The penetration penalty prevents implausible geometric intersections between the hand and the object. It penalizes points that violate a minimum separation margin τ_p:

$$\mathcal{L}_{\text{pen}} = \frac{1}{|\tilde{\mathcal{G}}_o|} \sum_{x_i \in \tilde{\mathcal{G}}_o} w_i \max(0, \tau_p - d_{i\to h})^2 + \frac{1}{|\tilde{\mathcal{G}}_h|} \sum_{x_j \in \tilde{\mathcal{G}}_h} \max(0, \tau_p - d_{j\to o})^2. \tag{4}$$

Crucially, w_i is a normalized weight proportional to the local deformation magnitude of the object. This deformation-aware reweighting encourages the optimization to prioritize physical collision resolution in highly deformed regions.

Furthermore, to encourage stable and realistic grasping, we apply an attractive contact loss $\mathcal{L}_{\text{attr}}$ exclusively to points within a proximity band τ_b:

$$\mathcal{L}_{\text{attr}} = \frac{1}{|\mathcal{C}_o|} \sum_{x_i \in \mathcal{C}_o} (d_{i\to h} - \tau_c)^2 + \frac{1}{|\mathcal{C}_h|} \sum_{x_j \in \mathcal{C}_h} (d_{j\to o} - \tau_c)^2, \tag{5}$$

where $\mathcal{C}_o = \{x_i \in \tilde{\mathcal{G}}_o \mid d_{i\to h} < \tau_b\}$ and $\mathcal{C}_h$ is defined similarly for the hand. Here, τ_c represents the ideal target distance for stable physical contact.

To encourage spatial smoothness of the force field and the material field, we apply total variation regularization on each spatial plane. Let μ denote one of the 2D feature planes and $\mu_{i,j}$ be the feature vector at location (i, j). The regularization term is defined as:

$$L_f = \sum_{i,j} |\mu_{i+1,j} - \mu_{i,j}|_2^2 + |\mu_{i,j+1} - \mu_{i,j}|_2^2 . \tag{6}$$

The total regularization term is defined as:

$$\mathcal{L}_{\text{reg}} = \mathcal{L}_{\text{tf}} + \mathcal{L}_{\text{mat}}. \tag{7}$$

4 Dataset

We construct a real-world multi-view dataset for bimanual interactions with deformable objects. The dataset contains 20 interaction sequences involving 10 everyday deformable objects and 1,000 selected frames. It is designed to capture articulated hand motion together with temporally varying object deformation under continuous bimanual interaction.

4.1 Capture and Interactions

The dataset is captured using six synchronized and calibrated RGB cameras at 30 FPS and 2160P resolution. The multi-view setup provides complementary observations under severe hand-object occlusion, which is common in bimanual manipulation, and enables the reconstruction of complete hand-object interaction states. The interactions cover stretching, squeezing, bending, pressing, compression, and pulling. During recording, the hands remain in continuous interaction with the objects, while the hand configurations, contact regions, and deformation patterns vary over time. This design captures both localized deformation around contact regions and larger-scale deformation distributed over the object.

4.2 Annotations and Intended Use

Each sequence provides synchronized multi-view RGB images, calibrated camera intrinsics and extrinsics, and hand-object segmentation masks. Using our multi-view reconstruction framework,

we further provide temporally aligned hand states, an object template, and per-frame deformable object geometry. These reconstructed annotations preserve not only the dynamic object shape but also the spatial and contact relationships between the articulated hands and the deforming object throughout the interaction.

Our dataset is designed with a similar goal to HMDO: providing temporally aligned hand states and dynamic deformable-object geometry to support research on deformable hand-object reconstruction. Since the reconstructed annotations are generated through our multi-view reconstruction pipeline, they are not considered independent ground truth for quantitatively evaluating our own reconstruction method. Instead, they provide supervision for downstream methods under more limited observation settings, such as monocular input. This enables the study of learning-based deformable hand-object reconstruction beyond the original multi-view capture scenario. Additional acquisition and annotation details are provided in the supplementary material.

5 EXPERIMENTS

5.1 DATASETS

ARCTIC dataset. We conduct our experiments on the ARCTIC (Fan et al., 2023) dataset, which provides sequences showing two hands skillfully manipulating various objects. The dataset provides accurate multi-view RGB images, hand and object meshes, and synchronized 3D annotations across sequences. We use only sequences categorized under the 'grab' action to ensure that the object remained rigid across the frames and selected 50 consecutive frames where the hand and object were clearly visible and in an interactive state.

HMDO dataset. We additionally evaluate our method on the HMDO (Xie et al., 2023) dataset, which focuses on markerless multi-view capture of hand–object interactions involving deformable objects. The dataset provides synchronized multi-view RGB videos, calibrated camera parameters, 3D hand poses, and high-quality deformable object surface reconstructions obtained through multi-view photometric observations. Unlike rigid-object datasets, HMDO includes interactions where objects exhibit significant elastic deformation, making it a challenging benchmark for evaluating non-rigid reconstruction.

For our experiments, we select sequences with predominantly continuous contact, and the interaction dynamics are clearly observable from multiple viewpoints. Each sequence consists of 30–50 consecutive frames, providing sufficient temporal continuity as well as clear visual evidence of both hand motion and interaction-induced object deformation.

5.2 METRICS

We evaluate our method using metrics that measure hand and object reconstruction accuracy. For hand reconstruction, we report the Mean Per-Joint Position Error (MPJPE, in millimeters) to evaluate the discrepancy between the predicted and ground-truth 3D hand joints. For object reconstruction, we use the Chamfer Distance (CD_o, in millimeters) between the predicted and ground-truth object surfaces, which reflects geometric reconstruction accuracy of deformable object surfaces. Following prior works (Aytekin et al., 2025; Liu et al., 2025; Wang et al., 2025; On et al., 2025), we additionally report the F-score, defined as the harmonic mean of precision and recall, computed between the predicted and ground-truth object surface point clouds at distance thresholds of 5 mm and 10 mm (F@5mm and F@10mm). We also report Mean Surface Penetration Depth (MSPD) and Penetrating Vertex Ratio (PVR) to evaluate local hand-object interaction consistency.

5.3 COMPARISON TO STATE-OF-THE-ART METHODS

Quantitative Results. Tab. 1 reports quantitative comparisons on the ARCTIC and HMDO datasets. Overall, our method achieves the best performance on most evaluation metrics across both datasets.

On hand reconstruction, we achieve lower MPJPE, indicating more accurate 3D hand pose estimation compared to IHOI (Ye et al., 2022), DiffHOI (Ye et al., 2023), and EasyHOI (Liu et al., 2025). This improvement is primarily attributed to our multi-view-based initialization, which provides more reliable interaction priors, as well as our image-driven optimization strategy that enforces

Table 1: Quantitative results on ARCTIC and HMDO datasets. The HMDO comparison is conducted on sequences $[1, 5, 7, 14, 15]$, the common subset for which all competing methods produce valid outputs.

	ARCTIC				HMDO			
Method	MPJPE↓	CD_o ↓	F@5mm↑	F@10mm↑	MPJPE↓	CD_o ↓	F@5mm↑	F@10mm↑
IHOI	36.12	6.28	8.79	20.01	19.25	5.50	10.52	25.15
DiffHOI	53.35	29.34	5.53	13.83	47.85	18.22	5.91	21.37
HOLD	**16.12**	3.38	34.65	66.62	**4.10**	10.84	9.37	29.38
EasyHOI	26.37	4.66	11.70	25.94	20.54	6.30	7.29	16.75
Ours	19.99	**1.11**	**52.14**	**90.75**	12.70	**2.82**	**20.20**	**45.12**

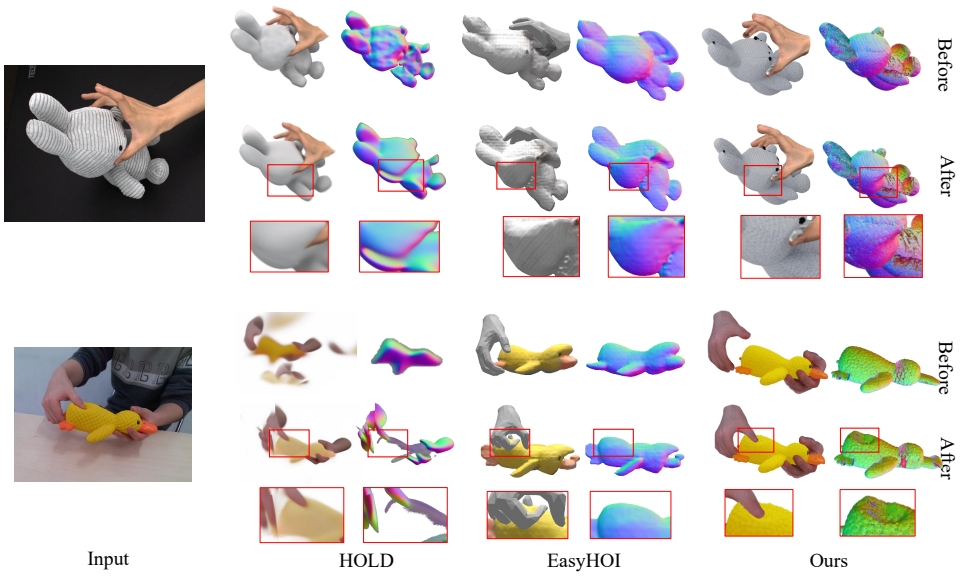


Figure 3: **Qualitative comparison.** We compare hand-object reconstruction results before and after interaction on HMDO (row 1) and our dataset (row 2). Our method better captures interaction-induced object deformation while maintaining hand-object alignment. Compared with template-based HOLD and EasyHOI, our method produces more accurate deformable object reconstruction under interaction.

hand-object consistency directly in the image space rather than relying solely on sparse keypoint supervision.

For object reconstruction, our approach significantly reduces the Chamfer Distance and achieves higher F-scores at both 5 mm and 10 mm thresholds. This improvement stems from two aspects. First, our SAM-based object initialization provides a stronger starting point compared to prior pipelines. Second, our deformation-aware optimization enables more accurate modeling of interaction-induced geometry changes, leading to improved surface reconstruction quality. Notably, the performance gains are more pronounced on the HMDO dataset, which contains deformable objects and more challenging interaction scenarios. This demonstrates the advantage of our framework in handling non-rigid object deformation, where methods relying on rigid or template-based assumptions become less effective.

Qualitative analysis. We compare the reconstructed hand-object interactions before and after deformation across different approaches. Our method accurately captures interaction-induced object deformation while preserving texture details and hand-object alignment. In particular, it recovers localized deformation around contact regions, such as squeezing and grasping, which is challenging for methods based on rigid or template-based object representations.

Compared with HOLD, our method models the dynamic deformation of objects during interaction, whereas HOLD relies on a fixed template representation and therefore cannot capture non-rigid object changes. EasyHOI can recover plausible interaction appearance, but its interaction alignment may contain artifacts under severe deformation. In contrast, our method produces more consistent hand-object configurations and deformation results.

Table 2: Ablation study on HMDO. All variants use identical sequences, initialization, and optimization settings.

	HMDO					
Method	MPJPE↓	CD_o ↓	F@5mm↑(%)	F@10mm↑(%)	MSPD↓(mm)	PVR↓(%)
w/o Rigid Opt.	24.28	2.95	19.57	46.38	5.83	2.45
w/o Deform. Opt.	23.86	2.16	18.83	54.80	2.41	0.82
w/o Interaction	19.37	2.12	21.54	56.98	2.14	0.68
Full	**19.33**	**2.09**	**21.58**	**57.96**	**1.75**	**0.28**

5.4 Ablation study

To comprehensively evaluate the effectiveness of our approach, we conduct three ablation studies to analyze the contributions of rigid Gaussian-based optimization, deformable physical optimization, and hand-object contact modeling. All experiments are performed on the HMDO dataset, as it contains deformable hand-object interactions required for evaluating the proposed deformation and interaction modeling components.

Specifically, we use HMDO sequences $[1, 3, 4, 5, 6, 7, 11, 14, 15, 17]$, which cover a broader range of deformable interaction patterns while allowing all ablation variants to be evaluated using the same initialization protocol. Quantitative results are summarized in Tab. 2.

Rigid Gaussian-based optimization. We evaluate the contribution of our rigid Gaussian-based initialization and alignment module. As a baseline, we replace Gaussian-based optimization with keypoint-based hand pose estimation and silhouette-based object pose optimization. Introducing rigid Gaussian optimization consistently improves MPJPE, CD_o, F@5mm, and F@10mm, demonstrating that the Gaussian representation with multi-view appearance supervision provides more accurate hand-object geometric alignment.

Deformable physical optimization. We assess the effect of deformable physical optimization by disabling the deformation module and directly using the rigid reconstruction results. While this configuration already achieves reasonable reconstruction accuracy, incorporating deformable optimization improves deformation-sensitive metrics, including F@5mm and F@10mm. This demonstrates that the proposed vision-physics optimization captures interaction-induced object deformation beyond rigid alignment.

Hand-object contact modeling. We evaluate the role of contact-aware constraints by removing the penetration and contact regularization terms between the hand and object. Without these constraints, global reconstruction metrics remain similar, while local contact consistency degrades significantly. MSPD increases from 1.75 mm to 2.14 mm and PVR increases from 0.28% to 0.68%. This demonstrates that the proposed contact modeling effectively reduces hand-object interpenetration and improves interaction consistency.

6 Conclusion

In this paper, we present a vision-physics inverse modeling framework for reconstructing bimanual dynamic deformable hand-object interactions. By coupling 3D Gaussian representations with differentiable MPM simulation, our method exploits temporal multi-view appearance changes to infer interaction-driven object deformation. Instead of relying on category-level templates or purely data-driven priors, we jointly optimize an effective time-invariant material field and a time-varying force field to recover detailed hand-object interaction states and deformation dynamics. We further introduce a real-world multi-view dataset containing diverse bimanual interactions with deformable objects. Extensive experiments demonstrate improvements over existing approaches in geometric reconstruction accuracy and hand-object interaction consistency.

Limitations and future work. Our framework has several limitations. First, it relies on dense multi-view observations and computationally intensive inverse optimization, motivating future extensions to sparse-view and monocular settings. Second, it currently models only elastic deformation, leaving more complex behaviors, such as plasticity and viscoelasticity, for future work. Finally, incorporating explicit frictional contact and tactile feedback may enable more detailed reconstruction of manipulation dynamics.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, proofreading, and improving the clarity and organization of the manuscript. These tools were used to assist with refining the presentation of ideas, improving readability, and identifying potential ambiguities in the written text.

We have not used generative AI tools for generating research results, designing experiments, developing the proposed methodology, conducting scientific analyses, or producing claims and conclusions presented in this paper. All technical contributions, experimental results, and scientific interpretations were developed and verified by the authors.

All AI-assisted modifications were carefully reviewed by the authors. We manually checked the correctness, consistency, and originality of the revised text, and verified that all technical statements accurately reflect our proposed method and experimental findings. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI tools.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of our proposed vision-physics inverse modeling framework in the main paper and supplementary material. The main paper presents the overall formulation, optimization pipeline, and key components, including Gaussian-based hand-object representation, differentiable MPM-based deformation modeling, and interaction-aware optimization. Additional implementation details, parameter settings, and derivations are provided in the supplementary material.

For experimental reproducibility, we describe the dataset acquisition protocol, evaluation metrics, and experimental settings in the main paper and supplementary material. In particular, we provide details of our real-world multi-view deformable hand-object interaction dataset, including data composition, annotation generation, and intended usage. The supplementary material further includes additional implementation details and qualitative/quantitative evaluation results.

We will release the source code, dataset, and trained models upon publication to facilitate future research and independent verification of our results.

REFERENCES

Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. *arXiv preprint arXiv:2508.18213*, 2025.

Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6121–6131, 2020.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7061–7071, 2025.

Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pp. 361–378. Springer, 2020.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12417–12426, 2021.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9044–9053, 2021.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European conference on computer vision*, pp. 231–248. Springer, 2022.

Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12890–12900, 2023.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. Hort: Monocular hand-held objects reconstruction with transformers. *arXiv preprint arXiv:2503.21313*, 2025.

Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4769–4778, 2023.

Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, et al. Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In *European Conference on Computer Vision*, pp. 284–303. Springer, 2024.

Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. 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.

Tao Du, Kui Wu, Pingchuan Ma, Sebastien Wah, Andrew Spielberg, Daniela Rus, and Wojciech Matusik. Diffpd: Differentiable projective dynamics. *ACM Transactions on Graphics (ToG)*, 41 (2):1–21, 2021.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 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.

Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *European Conference on Computer Vision*, pp. 428–448. Springer, 2024a.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. 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, 2024b.

Bowen Fu, Gu Wang, Chenyangguang Zhang, Yan Di, Ziqin Huang, Zhiying Leng, Fabian Manhardt, Xiangyang Ji, and Federico Tombari. D-sco: Dual-stream conditional diffusion for monocular hand-held object reconstruction. In *European Conference on Computer Vision*, pp. 376–394. Springer, 2024.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17461–17474, 2025.

Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. *CoRR*, 2019.

Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. Physical interaction: Reconstructing hand-object interactions with physics. In *SIGGRAPH Asia 2022 conference papers*, pp. 1–9, 2022.

Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, 2022.

Tianyu Huang, Haoze Zhang, Yihan Zeng, Zhilu Zhang, Hui Li, Wangmeng Zuo, and Rynson WH Lau. Dreamphysics: Learning physics-based 3d dynamics with video diffusion priors. *arXiv preprint arXiv:2406.01476*, 2024.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11107–11116, 2021.

Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7219–7230. IEEE, 2025a.

Shijian Jiang, Qi Ye, Rengan Xie, Yuchi Huo, and Jiming Chen. Hand-held object reconstruction from rgb video with dynamic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12220–12230, 2025b.

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

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. Mv-sam3d: Adaptive multi-view fusion for layout-aware 3d generation. *arXiv preprint arXiv:2603.11633*, 2026.

Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, and Cewu Lu. Chord: Category-level hand-held object reconstruction via shape deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9444–9454, 2023.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Isabella Liu, Hao Su, and Xiaolong Wang. Dynamic gaussians mesh: Consistent mesh reconstruction from dynamic scenes. *arXiv preprint arXiv:2404.12379*, 2024a.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. 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.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. pp. 21740–21751, 2024b.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 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.

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. 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.

Yik Lung Pang, Changjae Oh, and Andrea Cavallaro. Sparse multi-view hand-object reconstruction for unseen environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 803–810, 2024.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5865–5874, 2021a.

Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021b.

Chandradeep Pokhariya, Ishaan Nikhil Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. 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.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12242–12254, 2025.

Aditya Prakash, Matthew Chang, Matthew Jin, Ruisen Tu, and Saurabh Gupta. 3d reconstruction of objects in hands without real world 3d supervision. In *European Conference on Computer Vision*, pp. 126–145. Springer, 2024.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. 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.

Haozhe Qi, Chen Zhao, Mathieu Salzmann, and Alexander Mathis. 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.

Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15100–15111, 2023.

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

Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, and Jianxin Liao. Decoupled iterative refinement framework for interacting hands reconstruction from a single rgb image. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8014–8025, 2023.

Boxiang Rong, Artur Grigorev, Wenbo Wang, Michael J Black, Bernhard Thomaszewski, Christina Tsalicoglou, and Otmar Hilliges. Gaussian garments: Reconstructing simulation-ready clothing with photorealistic appearance from multi-view video. In *2025 International Conference on 3D Vision (3DV)*, pp. 1054–1063. IEEE, 2025.

Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. Showme: Benchmarking object-agnostic hand-object 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1935–1944, 2023.

Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Jean-Sébastien Franco, and Grégory Rogez. Host3r: Keypoint-free hand-object 3d reconstruction from rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7204–7213, 2025.

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pp. 581–600. Springer, 2020.

Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. Thor-net: End-to-end graformer-based realistic two hands and object reconstruction with self-supervision. *arXiv e-prints*, pp. arXiv–2210, 2022.

Hao Tian, Chenyangguang Zhang, Rui Liu, Wen Shen, and Xiaolin Qin. Interaction-aware 4d gaussian splatting for dynamic hand-object interaction reconstruction. 2025.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. Ho-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. *CoRR*, abs/2406.06843, 2024. doi: 10.48550/ARXIV.2406.06843. URL https://doi.org/10.48550/arXiv.2406.06843.

Rong Wang, Wei Mao, and Hongdong Li. Interacting hand-object pose estimation via dense mutual attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5735–5745, 2023.

Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. 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*, pp. 5957–5968, 2025.

Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3d from images and videos. *arXiv preprint arXiv:2404.06507*, 2024.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

Wei Xie, Zhipeng Yu, Zimeng Zhao, Binghui Zuo, and Yangang Wang. Hmdo: Markerless multi-view hand manipulation capture with deformable objects. *Graphical Models*, 127:101178, 2023.

Qingshan Xu, Jiao Liu, Shangshu Yu, Yuxuan Wang, Yuan Zhou, Junbao Zhou, Jiequan Cui, Yew-Soon Ong, and Hanwang Zhang. Neuspring: Neural spring fields for reconstruction and simulation of deformable objects from videos. *arXiv preprint arXiv:2511.08310*, 2025.

Jie Yang, Bingliang Li, Fengyu Yang, Ailing Zeng, Lei Zhang, and Ruimao Zhang. Boosting human-object interaction detection with text-to-image diffusion model. *arXiv preprint arXiv:2305.12252*, 2023a.

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20953–20962, 2022.

Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, and Cewu Lu. Poem: reconstructing hand in a point embedded multi-view stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21108–21117, 2023b.

Yuxin Yao, Siyu Ren, Junhui Hou, Zhi Deng, Juyong Zhang, and Wenping Wang. Dynosurf: Neural deformation-based temporally consistent dynamic surface reconstruction. In *European Conference on Computer Vision*, pp. 271–288. Springer, 2024.

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. 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.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 19717–19728, 2023.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. 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*, pp. 1911–1920, 2024.

Zhenjun Yu, Wenqiang Xu, Pengfei Xie, Yutong Li, Brian W Anthony, Zhuorui Zhang, and Cewu Lu. Dynamic reconstruction of hand-object interaction with distributed force-aware contact representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8590–8599, 2025.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 445–456, 2024.

Chenyangguang Zhang, Yan Di, Ruida Zhang, Guangyao Zhai, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ddf-ho: Hand-held object reconstruction via conditional directed distance field. *Advances in Neural Information Processing Systems*, 36:56871–56884, 2023.

Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, Fabian Manhardt, Bowen Fu, Federico Tombari, and Xiangyang Ji. Moho: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9992–10002, 2024a.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024b.

Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T Freeman. Physdreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*, pp. 388–406. Springer, 2024c.

Zhongqun Zhang, Jifei Song, Eduardo Pérez-Pellitero, Yiren Zhou, Hyung Jin Chang, and Aleš Leonardis. Ncrf: neural contact radiance fields for free-viewpoint rendering of hand-object interaction. In *2024 International Conference on 3D Vision (3DV)*, pp. 387–396. IEEE, 2024d.

Zehao Zhu, Jiashun Wang, Yuzhe Qin, Deqing Sun, Varun Jampani, and Xiaolong Wang. Contactart: Learning 3d interaction priors for category-level articulated object and hand poses estimation. In *2024 International Conference on 3D Vision (3DV)*, pp. 201–212. IEEE, 2024.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 813–822, 2019.

Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9054–9064, 2023.

A Technical appendices and supplementary material

A.1 Simulation of elastic materials

We focus on elastic objects, whose deformation behavior is governed by their material properties. The Cauchy stress is computed as:

$$\sigma = \frac{1}{\det(F)} \frac{\partial \psi}{\partial F} F^T, \tag{8}$$

where F is the deformation gradient and $\psi(F)$ is the strain energy density function Xie et al. (2024).

We adopt a Neo-Hookean formulation:

$$\psi(F) = \mu \sum_{i=1}^{d} (\sigma^i - 1)^2 + \frac{\lambda}{2} (\det(F) - 1)^2, \tag{9}$$

where σ^i are the singular values of F, and μ, λ are the Lamé parameters related to Young's modulus E and Poisson's ratio ν:

$$\mu = \frac{E}{2(1+\nu)}, \quad \lambda = \frac{E\nu}{(1+\nu)(1-2\nu)}. \tag{10}$$

Following observations in prior work Zhang et al. (2024c), optimizing Poisson's ratio provides limited benefit. Therefore, we fix ν and only optimize Young's modulus E.

Within the MPM, the object is discretized into:

Lagrangian points: carrying material properties (density ρ, Young's modulus E, Poisson's ratio ν) and state variables (position, velocity, deformation gradient F)

Eulerian grid: used to compute spatial derivatives (e.g., stress and forces) and perform time integration

External forces are optimized as effective driving fields whose spatial support adapts to different interaction patterns. This interaction governs the update of point velocities, which are subsequently transferred to the grid for integration. The resulting motion is mapped back to the Gaussian representation, producing physically consistent non-rigid deformation.

A.2 Hand Gaussian Initialization

We initialize the hand Gaussian representation from the MANO hand model. Specifically, we extract the linear blend skinning (LBS) weights $W \in \mathbb{R}^K$ from MANO, where $W^k(v)$ represents the influence of the k-th joint on the hand vertex v, and K denotes the number of hand joints.

The position of the i-th hand vertex in the posed space is computed by applying the LBS transformation:

$$\mu_i^p = \sum_{k=1}^{K} W^k(\mu_i^c)(\Phi_k \mu_i^c + \Gamma_k), \tag{11}$$

where Φ_k and Γ_k represent the rotation and translation of the k-th joint, which are computed from the MANO pose parameters θ_t and shape parameters β. The transformed hand vertices are used to initialize the hand Gaussian means.

During optimization, the Gaussian topology of the hand remains fixed. We optimize the hand pose sequence $\{\theta_t\}$ and per-Gaussian appearance parameters c_h using multi-view photometric supervision. The rendered hand appearance is supervised by a combination of L1 and SSIM losses.

A.3 Deformed 3DGS to Mesh Conversion

To support downstream applications and evaluate geometric accuracy, we extract explicit triangular meshes from our deformed 3D Gaussians at each frame. First, we filter out transparent floating artifacts by discarding Gaussians below a predefined opacity threshold, treating the remaining Gaussian centers as a dense point cloud. Since the Gaussians naturally flatten and align with the underlying object surface during the multi-view optimization process, we can analytically estimate their surface normals by extracting the direction of their shortest scaling axis. Given this oriented point cloud, we apply the Screened Poisson Surface Reconstruction (SPSR) algorithm to generate a smooth and watertight object mesh. Finally, the vertex colors are directly queried from the Spherical Harmonics (SH) coefficients of the nearest Gaussians, yielding a high-fidelity textured mesh sequence.

A.4 Comparison Experiments on Non-Rigid Object Reconstruction

To assess whether the proposed vision-physics formulation improves deformable-object reconstruction beyond conventional HOI baselines, we compare our method with PhysTwin, a closely related approach that reconstructs deformable objects through differentiable physical simulation. Although the two methods target different task settings, object geometry is a common output and therefore provides a direct basis for comparison. This experiment isolates the quality of non-rigid reconstruction and evaluates whether jointly modeling articulated hands, contact configurations, and interaction-driven deformation provides additional benefits under severe hand-object occlusion.

PhysTwin (Jiang et al., 2025a) represents deformable objects using a discrete spring-mass model. In contrast, our method jointly recovers time-varying interaction-driven loads, spatially varying material responses, and the resulting deformation process through differentiable Gaussian-MPM simulation. Consequently, our formulation explains not only how an object deforms, but also where

Table 3: Comparison with PhysTwin on HMDO. PhysTwin uses depth estimates from Depth Anything 3 (DA3).

Method	Reconstructed state	$CD_o \downarrow$ (mm)	F@5mm↑ (%)	F@10mm↑ (%)
PhysTwin + DA3	Object only	3.52	18.26	32.21
Ours	Hands + deformable object	**2.82**	**20.20**	**45.12**

the interaction induces deformation, how this driving effect evolves over time, and why different regions respond differently. Although PhysTwin estimates physical parameters for object dynamics, it does not jointly couple them with articulated hand poses and hand-object contact states. Moreover, our Gaussian particles provide a shared representation for both rendering and physical evolution, directly connecting visual evidence, contact constraints, and dynamic deformation.

We further compare the reconstructed object geometry of the two methods on HMDO. Since HMDO does not provide measured depth, we run PhysTwin using monocular depth estimates produced by Depth Anything 3 (DA3) (Lin et al., 2025). As shown in Tab. 3, our method achieves a lower Chamfer Distance and higher F-scores while additionally reconstructing the articulated hands and their interaction with the deformable object.

Under severe hand-induced occlusion, PhysTwin + DA3 is less accurate because its object-only formulation neither explicitly models the occluding hands nor exploits hand-object contact to constrain the hidden object regions. These results highlight our two main distinctions from PhysTwin: reconstructing the complete deformable hand-object manipulation state and introducing a field-driven Gaussian-MPM formulation.

A.5 Evaluation Protocol and Dataset Statistics

All metrics are computed on every evaluated frame, averaged within each sequence, and then averaged across sequences so that each sequence contributes equally to the final result. For ARCTIC, we evaluate 8 sequences comprising 240 frames and 5 objects. For HMDO, we evaluate 5 sequences comprising 150 frames and 5 objects. Our collected dataset contains 20 sequences, 10 objects, and 1,000 selected frames.

For the ablation study in Tab. 2, all variants are evaluated using identical sequences and frames, with the same initialization and optimization settings, thereby forming a controlled paired comparison. Removing the interaction regularization consistently degrades both hand and object metrics. The absolute changes in MPJPE and CD_o are modest because these metrics average errors over the entire hand and object, whereas the interaction terms act locally near contact regions, diluting their improvements in the global averages.

A.6 Physical and Geometric Consistency

Our reconstruction combines complementary physical and geometric constraints. The differentiable MPM simulation governs the non-rigid evolution of the object under the estimated force and material fields, constraining the recovered deformation to be physically plausible under the adopted elastic constitutive model. In parallel, the bidirectional contact objective penalizes implausible hand-object intersections and separation, thereby improving the geometric consistency of the reconstructed interaction. Because the recovered physical fields are effective latent simulator parameters rather than direct measurements, we assess their validity through the resulting deformation geometry and hand-object contact consistency.

Tab. 4 quantifies the contribution of the interaction constraints using Mean Surface Penetration Depth (MSPD) and Penetrating Vertex Ratio (PVR). Compared with the variant without interaction regularization, the full model reduces MSPD from 2.14 mm to 1.75 mm and PVR from 0.68% to 0.28%, corresponding to relative reductions of 18.2% and 58.8%, respectively. These results demonstrate that the interaction constraints substantially improve local geometric and contact consistency.

Table 4: Effect of the interaction constraints on hand-object geometric consistency.

Method	MSPD↓ (mm)	PVR↓ (%)
w/o Interaction	2.14	0.68
Full	**1.75**	**0.28**

Table 5: Utility of our dataset for downstream monocular hand-object reconstruction. "IHOI w/ our train set" denotes IHOI trained with our dataset. The HMDO evaluation uses the same sequences as the ablation study: $[1, 3, 4, 5, 6, 7, 11, 14, 15, 17]$.

Evaluation dataset	Method	MPJPE↓ (mm)	CD_o ↓ (mm)	F@5mm↑ (%)	F@10mm↑ (%)
Our test set	IHOI	18.72	8.33	9.46	24.05
Our test set	IHOI w/ our train set	**17.34**	**3.38**	**18.85**	**40.76**
HMDO	IHOI	25.32	10.87	8.33	22.16
HMDO	IHOI w/ our train set	**18.10**	**3.92**	**15.86**	**36.71**

A.7 Purpose and Utility of Our Dataset

Our dataset is constructed by fitting the proposed reconstruction framework to calibrated multi-view observations. Similar to HMDO, it is intended to provide a challenging benchmark and supervision resource for downstream monocular hand-object reconstruction, particularly for dynamic interactions involving deformable objects. Because the dynamic geometries are generated by our multi-view reconstruction pipeline, they can be used to evaluate downstream monocular methods, but do not constitute independent ground truth for quantitatively evaluating our own reconstruction framework.

Qualitative results on our dataset are presented in the second row of Fig. 3 and in the sixth example of the accompanying demo video. To further assess its downstream utility, we train IHOI with our training set and evaluate it on both our test set and HMDO. As shown in Tab. 5, training with our dataset improves all metrics in both the in-domain and cross-dataset settings.

The consistent improvements on our test set and HMDO indicate that the reconstructed annotations provide effective supervision for downstream hand-object reconstruction and improve cross-dataset generalization to deformable interactions.

A.8 Force field representation ablation

The external force field is a key component for modeling interaction-driven deformation. Instead of assigning independent forces to each object particle, we represent the force field using a continuous triplane representation. This design introduces a compact spatial prior and encourages smooth force variations among neighboring particles, while avoiding the large number of independent variables required by particle-wise optimization.

To analyze the effect of the force representation, we compare the proposed triplane-based force field with a particle-wise force optimization strategy. As shown in Fig. 4, both representations can produce deformation that matches the observed visual evidence. However, the particle-wise representation tends to generate scattered force distributions, especially around the interaction region, due to the lack of spatial continuity constraints. In contrast, the triplane representation produces smoother and more spatially coherent force patterns concentrated around the hand-object contact area, which better reflects the continuous nature of interaction forces.

For localized deformation, such as pressing, we constrain force optimization to particles within a 2 cm neighborhood of the hand-object contact region. For large-scale deformation, including stretching, compression, and bending, the force field is optimized over the entire object to capture distributed interaction effects.

A.9 Implementation details

Our framework is implemented using PyTorch and CUDA-based differentiable rendering and simulation modules. All experiments are conducted on a single NVIDIA RTX 4090D GPU with 48GB

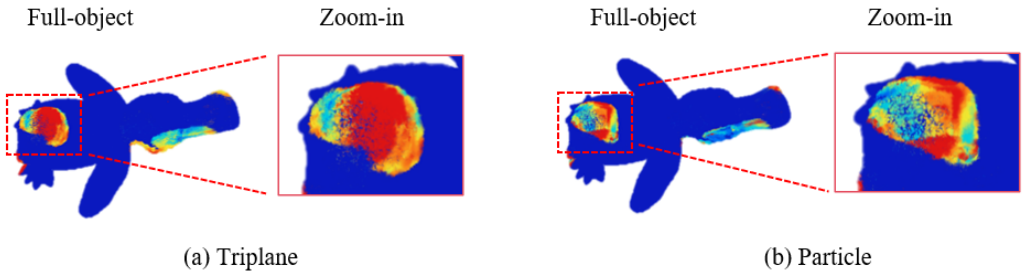


Figure 4: **Ablation of triplane force feild**

memory. For optimization, we use the Adam optimizer for hand pose optimization, rigid object pose optimization, and joint hand-object Gaussian optimization. The deformable physics stage propagates gradients through a differentiable MPM simulation pipeline.

Each optimization is performed on a single interaction sequence in a per-sequence optimization manner. The learning rate for rigid optimization is set to 1×10^{-3} and decayed exponentially during training, while the deformable optimization stage uses a smaller learning rate of 5×10^{-4} for stable physics-aware convergence. The Gaussian rendering resolution is set to 2160×3840. We initialize the object Gaussian representation from the object Gaussians reconstructed by MVSAM3D and initialize hand Gaussians from the MANO template. We perform two rounds of densification on the MANO template to meet our sampling requirements for hand vertices. Each sequence is optimized for 80 iterations during hand coloring, 400 iterations during rigid reconstruction, and 200 iterations during deformable optimization. For differentiable physics simulation, we adopt a Neo-Hookean constitutive model with fixed Poisson's ratio $\nu = 0.4$. The MPM simulation uses a grid resolution of 32 and a timestep of $1e-4$. We run 20 MPM steps for each frame. Gravity is disabled during optimization to focus on interaction-induced deformation. For the proposed contact loss, the penetration threshold τ_p is set to 5 mm, while the attraction bandwidth τ_b is set to 10 mm.

A.10 Additional Dataset Details

Capture setup. We use six synchronized RGB cameras arranged in a hemispherical configuration around the interaction space. All cameras are temporally synchronized and geometrically calibrated. We maintain stable illumination during recording to reduce photometric inconsistencies across views. The multi-view configuration provides complementary observations when the object is partially occluded by one or both hands, which is common during bimanual manipulation.

Object and interaction selection. The dataset includes deformable everyday objects with different shapes and deformation characteristics, including plush toys, plastic bottles, paper cups, seat cushions, rubber toys, and slippers. We design the interactions to cover both localized and distributed deformation. Pressing and squeezing mainly produce deformation around the contact regions, whereas stretching, compression, and bending affect larger parts of the object. From the captured sequences, we select 1,000 representative frames in which both the hand-object interaction and the resulting deformation are clearly observable.

Reconstructed annotations. In addition to the captured RGB images, calibrated camera parameters, and segmentation masks, we use our reconstruction framework to generate temporally aligned hand pose parameters, the object base model, and a deformable object mesh for each frame. We also provide the time-invariant material field and time-varying force field estimated by the inverse-physics optimization.

The reconstructed physical fields are effective latent simulator parameters used to explain the observed deformation rather than direct measurements of absolute material properties or interaction forces. Similarly, the reconstructed hand states and dynamic object geometry are method-generated annotations rather than independent ground-truth measurements. Their primary role is to provide

dynamic deformable hand-object geometry and supervision for downstream reconstruction methods.