CHOIR: Contact-aware 4D Hand-Object Interaction Reconstruction

HAO XU, The Chinese University of Hong Kong, China
YILIN LIU, University College London, United Kingdom
YINQIAO WANG, The Chinese University of Hong Kong, China
CHI-WING FU, The Chinese University of Hong Kong, China
NILOY J. MITRA, University College London, Adobe Research, United Kingdom

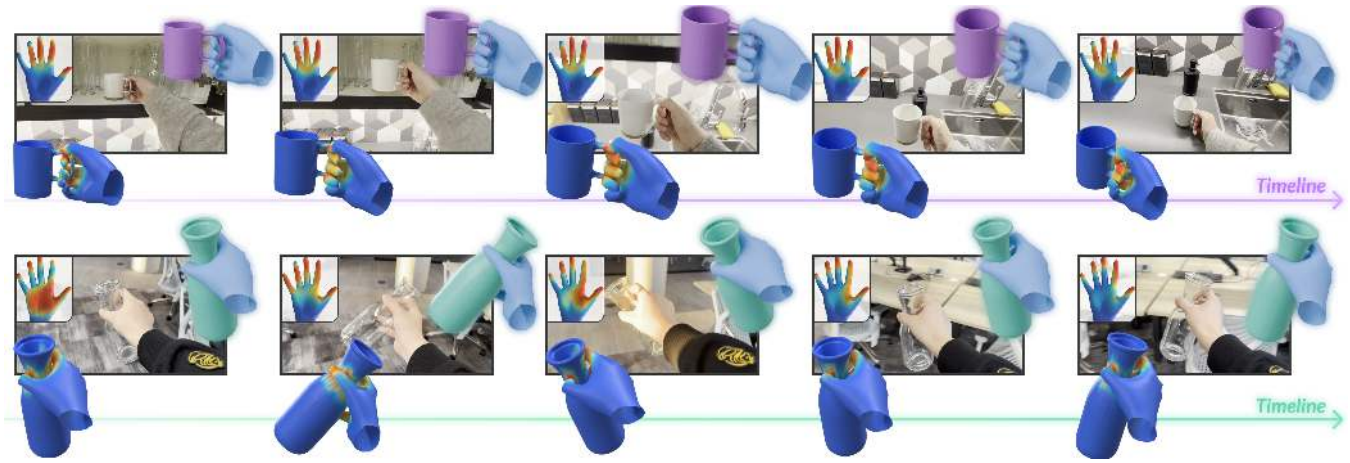

Fig. 1. From a monocular RGB video, CHOIR reconstructs 4D hand–object interactions (HOI) (including 3D hand motion, 3D object shape, 6D pose trajectory, and contact evidence) across open-world in-the-wild scenes. Heatmaps reveal the HOI contact information (see supplementary for videos).

We ask whether everyday open-world monocular videos can be turned into reusable 4D interaction primitives: articulated hand motion, object shape with 6D pose over time, and the when/where of contact. Such a capability would enable scalable mining of real interactions and, beyond reconstruction, support scene-aware synthesis and planning. However, reconstructing hand-object interaction (HOI) from challenging monocular videos remains difficult: methods often assume known objects or curated scenes, and separately estimated hands and objects easily become misaligned under clutter, occlusion, and unseen object geometries. Targeting this setting, we present CHOIR, a Contact-aware HOI Reconstruction framework for a monocular camera, using contact as an explicit coupling signal between hands and objects. CHOIR first initializes a coarse, contact-agnostic 4D HOI sequence from open-world visual priors. It then introduces a *generative HOI spatial rectification* module to predict ray-depth corrections and rectify hand-object relative placement, then derive initial per-frame contact correspondences on the rectified geometry. Last, a contact-aware joint optimization with dynamically updated contact constraints enforces geometric, temporal, and contact consistency. Experiments on controlled and challenging videos show that CHOIR improves object reconstruction, physical plausibility, and temporal consistency over state-of-the-art methods.

1 Introduction

> *"The world can only be grasped by action, not by contemplation. The hand is more important than the eye … The hand is the cutting edge of the mind."*
>
> — Jacob Bronowski, *The Ascent of Man* (1973)

Bronowski's observation frames hand-object interaction (HOI) through contact, the point at which perception becomes action. Inspired by this view, we aim to recover the underlying 4D interaction from a monocular RGB video: articulated hand motion, object shape and 6D pose over time, and when/where contact occurs. We focus on open-world manipulation videos in which the object and surrounding clutter are not known in advance, including challenging contact-bearing interaction phases with frequent occlusion and monocular depth ambiguity. If successful, such reconstruction would turn everyday videos into reusable interaction traces for mining, understanding, and scene-aware synthesis.

Despite rapid progress, current HOI pipelines still struggle outside curated settings. Many methods [Liu et al. 2021, 2025; Wang et al. 2025] are developed and validated under restrictive assumptions, such as known object models or categories, limited clutter, mild occlusion, or carefully controlled capture. Also, they often estimate hands and objects with weak coupling, without explicit constraints such as contact throughout reconstruction. As a result, performance degrades when the manipulated object is unseen, partially occluded, or visually ambiguous, which is precisely the regime of casual monocular videos such as Epic-Kitchens [Damen et al. 2018] and TASTE-Rob [Zhao et al. 2025]. Synthetic HOI data [Hasson et al. 2019; Wang et al. 2022] and laboratory captures [Chao et al. 2021; Hampali et al. 2020; Taheri et al. 2020; Yang et al. 2022; Zhan et al. 2024] provide valuable supervision, but diverse, physically grounded 4D interaction traces with object shape, object pose, hand motion, and explicit contact remain difficult to obtain at scale.

In such cases, contact provides a sparse but strong cue for hand-object coupling, yet it becomes reliable only after the relative hand-object placement has been corrected. Thus, the difficulty is not

merely to make each frame look plausible. A useful 4D HOI trace must simultaneously satisfy image evidence, metric object geometry, temporally coherent hand motion, and physically meaningful hand-object proximity. These requirements are tightly coupled: a visually plausible object mask may still imply the wrong metric scale, a smooth hand trajectory may float away from the object, and a contact cue measured before correcting relative depth can reinforce an incorrect surface correspondence.

This motivates a staged reconstruction strategy, in which we first establish a contact-agnostic 4D sequence (Stage 1: *Open-World HOI Analysis*), then we correct the interaction-frame hand-object placement (Stage 2: *Generative HOI Spatial Rectification*) and further stabilize the coupled motion over time (Stage 3: *Contact-Aware Optimization*); see Figure 2. Separating these roles also makes the pipeline easier to diagnose: each stage has a clear input, output, and failure mode before the final joint refinement, rather than hiding scale, contact, and temporal errors in a big optimization.

We design CHOIR, a contact-aware method to recover a structured 4D HOI trace from a monocular RGB video, while keeping the hand-object coupling explicit. First, we extract 2D interaction cues and initialize a coarse, contact-agnostic 4D HOI sequence. The object is reconstructed from an anchor frame using SAM-3D-Objects [SAM 3D Team et al. 2025], follow-tracked bidirectionally with guarded pose acceptance, and refined by isolated fitting, so that the sequence is visually aligned before contact reasoning. In parallel, hand motion is initialized and temporally stabilized using recent hand reconstruction priors [Pavlakos et al. 2024; Yu et al. 2025].

Second, independently initialized hands and objects can still misalign due to wrong relative depth, causing proximity-based contact search to fail or attach the hand to an incorrect object surface. This makes contact construction a consequence of spatial correction, rather than a standalone contact-map prediction problem: before asking which surfaces touch, the hand and object must first be rectified for spatial consistency. To do so, our generative HOI spatial rectification module uses a flow-matching prior to predict ray-depth corrections for interaction frames, refining the hand-object geometry into a contact-plausible relation, while preserving image evidence. Initial contact correspondences are then read out from the rectified geometry as hand contact vertices and barycentric object anchors. These correspondences serve as initialization evidence for optimization, rather than final reconstruction outputs. Finally, a contact-aware joint optimization starts from this rectified initialization, periodically rebuilds a soft contact cache from the current geometry, and combines contact, penetration, silhouette, anchor, and temporal losses to refine the full 4D sequence.

We evaluate CHOIR on controlled and challenging monocular videos. On HO3D [Hampali et al. 2020], CHOIR improves object reconstruction and hand-object alignment over state-of-the-art baselines [Fan et al. 2024; Liu et al. 2025; Wang et al. 2025]. For in-the-wild evaluation, we report reference-free physical and temporal metrics on a 100-video benchmark from TASTE-Rob [Zhao et al. 2025] and self-captured videos, with supplementary multi-view video comparisons for temporal assessment. These results show that our rectification and contact-aware optimization improve physical plausibility and stability while preserving image alignment. See Figure 1.

In summary, our contributions are threefold. First, we introduce an open-world monocular 4D HOI reconstruction method that extracts 2D interaction cues, performs guarded object tracking, and produces a coarse contact-agnostic sequence. Second, we propose generative HOI spatial rectification to predict ray-depth corrections before constructing initial per-frame barycentric contact correspondences. Third, we develop a contact-aware joint optimization with dynamically updated contact constraints and validate it on controlled and in-the-wild videos.

2 Related Work

2D Hand-Object Understanding. Reliable 2D spatial understanding is a prerequisite for 3D/4D hand-object reconstruction. Early works, *e.g.*, MediaPipe [Lugaresi et al. 2019] and 100DOH [Shan et al. 2020], established valuable benchmarks for 2D hand understanding, whereas recent large-scale approaches [Cheng et al. 2023; Potamias et al. 2025; Xu et al. 2022] further scaled up the performance. However, domain shifts in unconstrained videos [Zhao et al. 2025] with cluttered backgrounds remain difficult to resolve.

Monocular 3D/4D Hand Reconstruction. 3D hand reconstruction has evolved from single-image mesh recovery to 4D tracking. Early parametric methods that fit MANO [Romero et al. 2017] to 2D observations [Boukhayma et al. 2019] have been largely superseded by transformer-based architectures. Notably, HaMeR [Pavlakos et al. 2024] leverages large-scale Vision Transformers for robust single-image estimation, whereas WiLoR [Potamias et al. 2025] refines poses for better image alignment. Regarding videos, methods like SeqHAND [Yang et al. 2020] and Deformer [Fu et al. 2023] enforce temporal smoothness. Recent approaches such as Dyn-HaMR [Yu et al. 2025], HaWoR [Zhang et al. 2025b], and HaPTIC [Ye et al. 2026] advance 4D hand motion recovery by modeling camera dynamics and multi-frame priors. While these methods excel at isolated hand motion, they do not reconstruct manipulated object or model contact. In contrast, we tackle open-world HOI by jointly capturing both hand and object motions under explicit physical constraints.

Monocular 3D/4D Hand-Object Reconstruction. Recovering HOI is challenged by unknown object geometry, severe occlusions, and coupled dynamics. Early learning-based methods focused on monocular hand motion capture [Zhou et al. 2020] or aligning pre-scanned object templates [Hasson et al. 2019]. To move beyond fixed templates, subsequent works adopted neural implicit representations [Chen et al. 2022; Ye et al. 2022]. Regarding videos, HOLD [Fan et al. 2024] targets category-agnostic reconstruction over time, while DiffHOI [Ye et al. 2023] and G-HOP [Ye et al. 2024] utilize diffusion-based refinement. More recently, generative priors (*e.g.*, EasyHOI [Liu et al. 2025], MagicHOI [Wang et al. 2025]) synthesize plausible hand-object poses to handle occlusions. This direction also benefits from broader progress in 3D generative models [Hu et al. 2023b; Hui et al. 2022b; Liu et al. 2023], including controllable 3D and scene generation [Feng et al. 2025; Hu et al. 2023a, 2024, 2026; Hui et al. 2022a; Liu et al. 2026; Yan et al. 2026], 3D scene segmentation [Zhu et al. 2024a, 2025a,b], and geometric analysis [Du et al. 2025; Zhu et al. 2021, 2024b]. Conceptually related, BimArt [Zhang et al. 2025a] introduces a generative prior for HOI, but focuses on the forward *synthesis* of bimanual interactions.

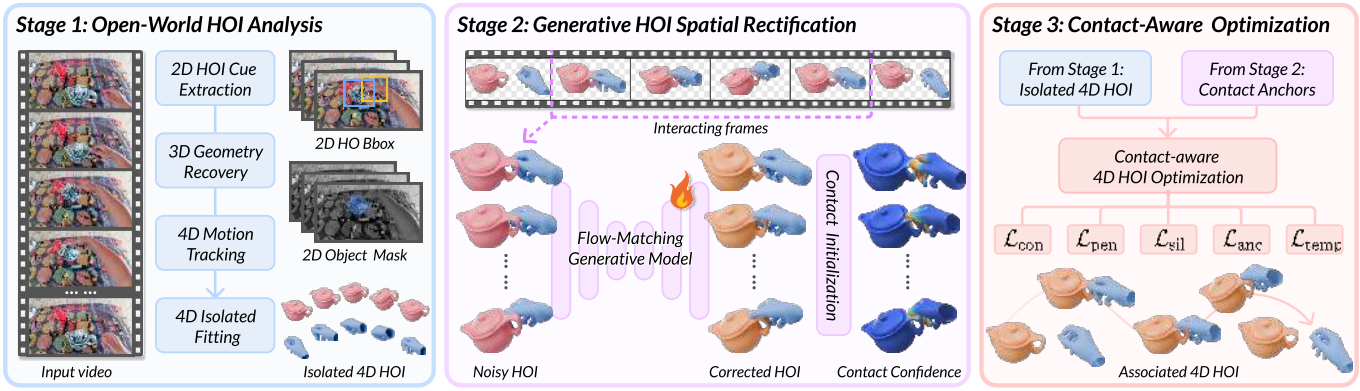

Fig. 2. Method overview. Stage 1: From a monocular video, we first obtain 2D HOI cues and initialize 3D hand/object reconstructions to form a coarse 4D HOI sequence. Stage 2: A flow-matching generative HOI spatial rectification module rectifies hand-object relative placement before initial contact correspondence construction. Stage 3: Hierarchical contact-aware optimization refines the sequence across phases, yielding physically consistent 4D hand–object interactions.

Despite these advances, recovering high-fidelity 4D HOI in an open-world setting remains non-trivial. First, while implicit representations, *e.g.*, HOLD, DiffHOI, and G-HOP, offer topological flexibility, preserving geometric details from monocular inputs remains ambiguous without explicit 3D priors. Second, generative approaches focus primarily on visual plausibility. EasyHOI [Liu et al. 2025] excels in single-view reconstruction but does not explicitly model temporal and contact coherence required for 4D motion, while MagicHOI [Wang et al. 2025] remains sensitive to initialization errors, occlusions, and contacts over occluded regions. In contrast, CHOIR targets automated 4D reconstruction with explicit contact correspondences and contact-aware joint optimization. Last, while optimization-based post-processing is highly effective for enforcing validity across diverse domains [Grady et al. 2021; Shen et al. 2025], achieving automated, physically-grounded 4D reconstruction across an interaction lifecycle remains largely unexplored.

3 Overview

Problem definition. Our goal is to recover a coherent 4D representation of hand-object interactions (HOI) from a monocular RGB video captured in an open-world setting. Specifically, the input video may capture either a complete atomic interaction, where the hand approaches a static object, grasps and manipulates it, and then returns the object to the scene (see Figure 3), or only the active object manipulation phase. The target output includes the hand motion, object shape, object 6D pose trajectory, and contact evidence for spatially and temporally consistent 4D HOI reconstruction. To make this goal tractable, we assume the object is rigid.

The task presents several challenges:

(i) *2D Perception Dilemma.* Specialized interaction detectors [Cheng et al. 2023; Leonardi et al. 2024; Shan et al. 2020] fall short of generalizing to unseen scenarios, while general-purpose models [Karaev et al. 2025; Potamias et al. 2025; Ravi et al. 2024] cannot readily comprehend HOI semantics out-of-the-box. There is no off-the-shelf pipeline to robustly recognize, segment, and track the hand and specific associated object during the interaction.

Fig. 3. Problem definition: our goal is to reconstruct 4D HOI from a monocular RGB video of a complete atomic interaction that includes (i) hand approaching a static object; (ii) grasping and manipulating the object; and (iii) putting the object back. Active-manipulation-only clips are also allowed.

(ii) *3D Spatial Ambiguity.* Monocular depth and scale ambiguities, especially under hand-object occlusions, make reliable spatial alignment difficult between the hand and the object. Such misalignment can make proximity-based contact search fail or attach the hand to an incorrect object surface.

(iii) *4D Physical Complexity.* Extending to 4D requires maintaining physically valid contact over the frames and phases where contact is present, while also preserving image alignment in non-contact frames. Since contact evidence is sparse and phase-dependent, temporal physical validity must be enforced together with penetration, silhouette, anchor, and motion priors; otherwise, the reconstruction is prone to interpenetration, floating artifacts, or temporal drift.

Method overview. Our key observation is that *contact* provides an explicit and transferable cue for hand-object coupling during interaction. Though monocular perception yields noisy hand and object estimates, generative HOI spatial rectification followed by contact-aware test-time optimization refines them into a physically consistent relation. CHOIR has three stages (Figure 2): (i) *Open-world HOI analysis* extracts 2D interaction cues, recovers hand/object geometry, tracks motion, and performs coarse isolated fitting to derive a contact-agnostic 4D HOI sequence (Section 4.1); (ii) *Generative HOI spatial rectification* predicts ray-depth corrections with a flow-matching prior to rectify the interaction-frame hand-object placement; per-frame contact correspondences are then read out as barycentric anchors on the rectified geometry (Section 4.2); and

(iii) *Contact-aware joint optimization* starts from the rectified initialization and contact evidence to jointly refine hand/object trajectories with dynamically rebuilt soft contact constraints together with silhouette, penetration, anchor, and temporal losses (Section 4.3).

4 Method

4.1 Stage 1: Open-World HOI Analysis

Given a monocular RGB video, Stage 1 extracts reliable image evidence and lifts it to an isolated, metrically plausible, contact-agnostic 4D HOI sequence (Figure 2). This sequence provides the coarse interaction geometry that Stage 2 rectifies before the subsequent contact-aware optimization in Stage 3. We summarize the main components below and defer details to the supplementary material.

Step 1: 2D interaction cue extraction. For each frame, we estimate 2D interaction cues: the hand bounding box, chirality, and 2D joints; the interacting-object bounding box; and modal/amodal hand-object masks. Generic foundation models [Cheng et al. 2023; Potamias et al. 2025] still suffer from chirality errors, missed detections, and weak object localization in out-of-distribution videos. We thus use a self-adaptation procedure to mine object pseudo labels around a motion-selected interaction frame, and utilize these labels to finetune WiLoR [Potamias et al. 2025] with an additional interacting-object box head, and adopt SAM 2 propagation [Ravi et al. 2024] along with amodal video segmentation [Chen et al. 2025] to obtain temporally complete masks. Please refer to the supplementary material for more details.

Step 2: 3D geometry recovery. HaMeR [Pavlakos et al. 2024] initializes per-frame MANO hand pose and camera-relative translation. SAM-3D-Objects [SAM 3D Team et al. 2025] reconstructs a canonical object mesh, metric scale, and initial 6D pose from an object anchor frame. This anchor frame defaults to the first frame but can be any frame with reliable object visibility, and its reconstruction provides the fixed object geometry used for follow tracking.

Step 3: 4D motion tracking. To lift the isolated 3D estimates into a coherent reconstruction-frame sequence, VIPE [Huang et al. 2025] provides camera estimates and Dyn-HaMR [Yu et al. 2025] stabilizes the initial MANO estimates into a 4D hand trajectory. For the object, we use SAM-3D-Objects as a guarded follow tracker: starting from the anchor-frame reconstruction, we freeze the shape, scale, and translation-scale latents and update only the rotation and translation pose latents, yielding bidirectional 6D object poses over time. Outlier poses are rejected, missing poses are interpolated, and the resulting dense object pose sequence initializes the coarse isolated fitting in Step 4. The strategies are detailed in the supplementary material.

Step 4: Coarse isolated sequence fitting. The tracked 4D sequence obtained above is still a coarse initialization only: object poses may drift under occlusion and hand estimates may contain 2D reprojection errors or implausible articulation. We thus run an isolated fitting step before HOI spatial rectification in Stage 1, since this step still relies on image evidence and generic priors, and its goal is to produce a temporally usable coarse 4D HOI sequence for Stage 2.

Regarding the hand, we optimize the MANO parameters with terms for preserving image alignment, monocular-depth consistency, plausible articulation, and temporal stability:

$$\mathcal{L}_h^{\text{iso}} = \mathcal{L}_{2D}^h + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}^h + \lambda_{\text{anat}}\mathcal{L}_{\text{anat}}^h + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}^h + \lambda_{\text{temp}}^h\mathcal{L}_{\text{temp}}^h, \quad (1)$$

where $\mathcal{L}_{2D}^h$ keeps the projected joints aligned with 2D evidence; $\mathcal{L}_{\text{depth}}^h$ stabilizes wrist depth using the median metric depth inside the hand mask; $\mathcal{L}_{\text{anat}}^h$ suppresses anatomically invalid finger poses; $\mathcal{L}_{\text{prior}}^h$ softly anchors the solution to the HaMeR initialization; and $\mathcal{L}_{\text{temp}}^h$ reduces frame-to-frame jitter. Regarding the object, direct mask-IoU optimization gives weak or unstable gradients when the rendered and target masks do not overlap. We thus use complementary silhouette terms with temporal and static consistency:

$$\mathcal{L}_o^{\text{iso}} = \lambda_{\text{rep}}\mathcal{L}_{\text{rep}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}} + \lambda_{\text{temp}}^o\mathcal{L}_{\text{temp}}^o + \lambda_{\text{stat}}\mathcal{L}_{\text{stat}}^o, \quad (2)$$

where $\mathcal{L}_{\text{rep}}$ penalizes rendered pixels that spill outside the amodal mask, preventing the rendered object from leaking into background regions; $\mathcal{L}_{\text{attr}}$ pulls projected vertices toward uncovered target-mask regions, providing useful gradients even before the masks overlap; $\mathcal{L}_{\text{temp}}^o$ smooths pose changes under occlusion; and $\mathcal{L}_{\text{stat}}^o$ ties static segments to a consistent object state. Compactly, with the rendered silhouette $\mathbf{M}$, amodal mask $\hat{\mathbf{M}}$, object vertices $\mathcal{V}$, camera projection $\Pi(\cdot)$, image domain Ω, sampled uncovered target-mask pixels $\mathcal{S}$,

$$\begin{aligned} \mathcal{L}_{\text{rep}} &= \frac{1}{|\Omega|}\sum_{\mathbf{p}\in\Omega}\left[\max\left(\mathbf{M}_\mathbf{p} - \hat{\mathbf{M}}_\mathbf{p}, 0\right)\right]^2, \\ \text{and } \mathcal{L}_{\text{attr}} &= \frac{1}{|\mathcal{S}|}\sum_{\mathbf{p}\in\mathcal{S}}\min_{\mathbf{v}\in\mathcal{V}}\|\mathbf{p} - \Pi(\mathbf{v})\|_2^2. \end{aligned} \quad (3)$$

After this step, we apply a ray-scale alignment that slides the object sequence along the camera ray, so that its interaction-depth statistics match the hand, preserving the image-space silhouette fit while improving the depth initialization for the Stage 2 rectification. Exact target-pixel sampling, EMA normalization, weights, and schedules are reported in the supplementary material.

4.2 Stage 2: Generative HOI Spatial Rectification

The Stage 1 sequence is visually aligned but not necessarily physically valid: the hand can hover, penetrate the object, or drift under occlusion. Stage 2 first rectifies the relative hand-object placement over the interaction range, moving the noisy hand estimate into a contact-plausible alignment before any contact anchors are constructed. It then reads out initial per-frame contact correspondences from the rectified geometry, represented by hand vertices and barycentric object anchors, rather than outputting a final optimized pose.

Modeling interaction inconsistency. We learn a conditional model of plausible grasps using large-scale synthetic grasping data with physical filtering. Starting from DexGraspNet [Wang et al. 2022], we simulate grasps over diverse object shapes of randomized scales and initial placements (table-top and mid-air). We then prune unstable grasps using a physics engine (namely, PyBullet [Coumans and Bai 2016]), retaining only the configurations that remain stable under gravity and perturbations, *e.g.*, shaking or perturbation forces. This yields a set of *contact-valid* target grasps.

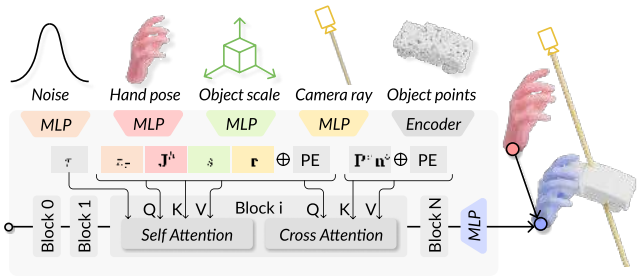

Fig. 4. Our flow-matching-based generative HOI spatial rectification.

To mimic monocular ambiguity, we generate a corresponding *noisy* source grasp for each simulated hand grasping pose by injecting *anisotropic ray-aligned* perturbations that emphasize depth uncertainty: large variance along the camera ray, mild in-plane noise, and anatomy-preserving perturbations in the MANO parameters. Each training pair thus consists of (i) a noisy hand estimate and viewing direction, (ii) the object geometry and pose context, and (iii) a physically valid target grasp with contact supervision. Examples of our paired data can be found in Figure 6.

Generative HOI spatial rectification. Given an object geometry and an initial (possibly misaligned) hand estimate, our goal is not to regress a single deterministic pose but to model the *distribution* of physically plausible corrections with contact consistency. We thus formulate rectification as a conditional generative process to predict a contact-consistent hand–object alignment under occlusion.

Concretely, we represent the object by surface points (with normals) and condition the model on the current 3D hand joints, viewing direction, and object scale. We parameterize the dominant monocular failure mode (*i.e.*, *depth ambiguity*) as a 1D offset along the camera ray, and learn a flow-matching-based model [Lipman et al. 2023] that maps a noisy depth offset to a rectified offset consistent with the object geometry, as Figure 4 shows. Let $\mathbf{P}^o$ denote sampled object surface points, $\mathbf{n}^o$ the associated normals, $\mathbf{J}^h$ the noisy hand joints, $\mathbf{r}$ the viewing ray, and s the object scale. The model predicts a scalar correction along viewing ray $\mathbf{r}$, because this depth direction captures the dominant monocular ambiguity while preserving the image-space hand configuration. With flow-matching model v_θ, time $\tau \in [0, 1]$, and state $z_\tau \in \mathbb{R}^1$, where z_τ is the noisy intermediate ray-offset state between the source offset and the rectified target offset, the rectification velocity is

$$\frac{\mathrm{d}z_\tau}{\mathrm{d}\tau} = v_\theta(z_\tau, \tau \mid \mathbf{J}^h, \mathbf{r}, s, \mathbf{P}^o, \mathbf{n}^o). \tag{4}$$

For each frame in the interaction range, sampling this conditional flow predicts the ray-depth correction that places the noisy hand estimate near a contact-plausible object surface while preserving its image-space evidence. Initial contact anchors are then computed from this rectified hand-object geometry. The Transformer tokenization, self-/cross-attention blocks, and AdaLN time conditioning are detailed in the supplementary material.

Contact correspondence establishment. For each interaction frame, the rectified hand estimate exposes a sparse contact structure for the contact-aware optimization in Section 4.3. To compute anchors on the corrected relative geometry, we align the per-frame object mesh to its canonical space via Procrustes registration and transform the rectified hand vertices accordingly. Within this canonical space, we densely sample the object surface and retrieve the K-nearest surface points for each candidate MANO contact vertex, where the candidates cover fingertips, finger pads, and palm regions. A valid correspondence is established only if it satisfies two geometric criteria: (i) the Euclidean distance falls below a proximity threshold, and (ii) the object-surface direction is compatible with the corresponding fingerpad normal, ensuring that the anchor lies on a physically plausible contact side. The geometrically validated surface points are then projected onto the object mesh and parameterized as barycentric coordinates (face_id, barycentric). These initial per-frame anchors provide contact evidence for Stage 3, where the soft contact cache is rebuilt and stabilized during the optimization.

4.3 Stage 3: Contact-aware HOI Optimization

Stage 1 provides a visually-aligned but not necessarily physically valid output, whereas Stage 2 provides a rectified interaction sequence with contact correspondences but not final poses. Stage 3 then jointly optimizes the hand and object trajectories from the rectified initialization, using the Stage 2 correspondences as contact evidence and image-alignment and temporal priors from Stage 1.

For each clip, we consider up to five phases–pre-static, approach, manipulation, release, and post-static–although many clips contain only a subset of them (fig. 3). Contact losses are active on frames where contact correspondences are available; silhouette, penetration, anchor, and temporal terms provide context across the whole sequence. Phase detection details, thresholds, and schedules are provided in the supplementary material. The contact-aware optimization minimizes five weighted objectives:

$$\mathcal{L}_{\text{HOI}} = \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{pen}}\mathcal{L}_{\text{pen}} + \lambda_{\text{sil}}\mathcal{L}_{\text{sil}} + \lambda_{\text{anc}}\mathcal{L}_{\text{anc}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}}. \tag{5}$$

$\mathcal{L}_{\text{con}}$ pulls active contact vertices toward object-surface anchors; $\mathcal{L}_{\text{pen}}$ penalizes interpenetration; $\mathcal{L}_{\text{sil}}$ aligns rendered silhouettes to amodal masks; $\mathcal{L}_{\text{anc}}$ collects object/hand anchors and reliable 2D/anatomical evidence; and $\mathcal{L}_{\text{temp}}$ enforces smooth motion.

Contact and penetration constraints. Starting from the Stage 2 rectified interaction frames, the optimizer periodically rebuilds a soft top-K correspondence cache from the current hand-object geometry. Each active MANO contact vertex is associated with object-surface anchors represented by face ids and barycentric coordinates. A representative term in $\mathcal{L}_{\text{con}}$ is the soft correspondence loss

$$\mathcal{L}_{\text{contact}} = \frac{\sum_{t,i} c_{t,i} \sum_{k=1}^{K} \omega_{t,i,k} \|\mathbf{v}^h_{t,i} - \mathbf{a}^o_{t,i,k}\|_2^2}{\sum_{t,i} c_{t,i}}, \tag{6}$$

where $\mathbf{v}^h_{t,i}$ is hand contact vertex i at frame t; $c_{t,i}$ is its contact confidence/gate, computed from the rebuilt cache using proximity, normal compatibility, and temporal memory; $\omega_{t,i,k}$ is a normalized softmax weight over the top-K candidate object anchors, computed from anchor distances; and $\mathbf{a}^o_{t,i,k}$ is the dynamically rebuilt barycentric object anchor. If no active contacts are available in a clip, we skip $\mathcal{L}_{\text{contact}}$. In parallel, a one-sided penetration term, evaluated over all optimized frames, pushes hand vertices that lie inside the object back toward the object surface. Let $\mathcal{V}^h = \{\mathbf{v}^h\}$ be the set of

hand mesh vertices, $d_o(\mathbf{v}^h)$ be the signed penetration depth that is positive inside the object and non-positive outside, and ϵ_{pen} be a small tolerance. We use

$$\mathcal{L}_{\text{pen}} = \frac{1}{|\mathcal{V}^h|} \sum_{\mathbf{v}^h} \max\left(d_o(\mathbf{v}^h) - \epsilon_{\text{pen}},\ 0\right). \tag{7}$$

This stage keeps the object pose trainable, where the silhouette and anchor groups prevent contact constraints from moving the object away from the image evidence.

Silhouette, anchor, and temporal stabilization. The remaining loss families stabilize the optimization when contact evidence is sparse or phase-dependent. We use the following compact definitions:

$$\begin{aligned} \mathcal{L}_{\text{sil}} &= \frac{1}{T} \sum_t \ell_{\text{mask}}\left(\mathcal{R}(\mathbf{O}_t), \hat{\mathbf{M}}_t^o\right), \\ \mathcal{L}_{\text{anc}} &= \mathcal{L}_{2D}^h + \mathcal{L}_{\text{anat}}^h + \mathcal{L}_{\text{pose}}^h + \mathcal{L}_{\text{pose}}^o, \\ \text{and } \mathcal{L}_{\text{temp}} &= \frac{1}{T} \sum_t \left(\|\Delta \mathbf{x}_t\|_2^2 + \|\Delta^2 \mathbf{x}_t\|_2^2\right), \end{aligned} \tag{8}$$

where T is the number of optimized frames; $\mathcal{R}(\mathbf{O}_t)$ is the rendered object silhouette from object state $\mathbf{O}_t$; $\hat{\mathbf{M}}_t^o$ is the amodal object mask; ℓ_{mask} is the silhouette discrepancy implemented as the MSE between the rendered and target amodal masks; $\mathcal{L}_{2D}^h$, $\mathcal{L}_{\text{anat}}^h$, $\mathcal{L}_{\text{pose}}^h$, $\mathcal{L}_{\text{pose}}^o$ represent anchor terms on 2D hand joint, 3D hand anatomy, hand pose, and object pose, respectively; and $\mathbf{x}_t$ collects the optimized hand pose, wrist motion, and object pose, with Δ and Δ^2 denoting the first- and second-order temporal finite differences.

Details on the contact-cache construction, loss components and weights, thresholds, and optimization schedules are provided in the supplementary material.

5 Experiments

Baselines. We compare CHOIR against five representative state-of-the-art template-free HOI reconstruction baselines: iHOI [Ye et al. 2022], DiffHOI [Ye et al. 2023], HOLD [Fan et al. 2024], EasyHOI [Liu et al. 2025], and MagicHOI [Wang et al. 2025].

Training dataset. To train our generative HOI spatial rectification module (Section 4.2), we build *GraspPair* from DexGraspNet [Wang et al. 2022], a large-scale synthetic dataset containing ~500k paired hand-object grasps over 5,355 object instances from 133 categories. Each object is paired with 200 physically valid grasp configurations on average, providing diverse compatible 3D hand poses. These pairs supervise the flow-matching model to learn plausible hand-object spatial relationships that generalize to unseen objects.

Evaluation benchmarks. We evaluate metric accuracy and open-world robustness using data from three sources. (i) *HO3D* [Hampali et al. 2020] provides lab-captured videos with ground-truth 3D annotations; following standard protocols [Fan et al. 2024; Wang et al. 2025], we use its 14-sequence subset for quantitative metric evaluation. (ii) *TASTE-Rob* [Zhao et al. 2025] provides diverse monocular sequences that feature rigid objects in cluttered backgrounds and long-tail scenarios, from which we select 70 challenging single-hand clips. (iii) *Self-captured videos* comprise 30 casually recorded clips of diverse objects, scenes, and motions to test unconstrained generalizability. To assess *in-the-wild* performance, where 3D ground truths are not available, we curate a comprehensive benchmark combining TASTE-Rob and self-captured videos. For each baseline, we compute metrics only on sequences for which at least one third of frames produce a reconstruction. As the baseline outputs contain outliers, we report mean, standard deviation, and median for the in-the-wild comparison. Details are provided in the supplementary material.

Table 1. Quantitative comparison on HO3D using ground-truth reconstruction metrics (top) and reference-free metrics (bottom). Best results are bold.

Method	CD	F5	F10	MPJPE	CD_h	RS
	[cm] ↓	[%] ↑	[%] ↑	[mm] ↓	[cm] ↓	↓
iHOI	2.37	35.78	62.11	27.75	25.45	0.17
DiffHOI	2.30	39.59	64.49	16.02	33.33	0.13
EasyHOI	1.86	46.10	70.92	16.69	19.55	0.28
HOLD	1.31	57.20	80.23	30.79	21.28	0.62
MagicHOI	0.87	69.72	92.15	**4.62**	**2.39**	0.11
Ours	**0.77**	**72.33**	**96.03**	5.55	2.41	**0.10**

Method	mIoU	Pen. Ratio	H-O Dist.	Acc_h	Acc_o
	[%] ↑	[%] ↓	[cm] ↓	[cm/fr^2] ↓	[cm/fr^2] ↓
HOLD	75.01	17.72	0.10	2.07	**1.03**
MagicHOI	70.57	14.11	0.36	9.51	7.51
Ours	**85.87**	**4.58**	**0.06**	**1.75**	1.08

Metrics. We evaluate hand pose, object shape, and hand–object alignment quality. Following [Fan et al. 2024; Wang et al. 2025; Ye et al. 2023], we report root-relative MPJPE (mm) for hand joints and Chamfer Distance (CD, cm) for object reconstruction. For the remaining HO3D object-centric metrics, we follow the MagicHOI evaluation protocol [Wang et al. 2025] to compute F-scores at 5 mm (F5) and 10 mm (F10) for local shape detail. For hand-relative alignment, we translate the predicted object mesh by the estimated hand-root, then compute hand-relative Chamfer distance (CD_h). Lastly, we report Relative Scale (RS), where s_{ICP} is the ICP scale to align the predicted mesh to ground truth. $\text{RS} = \frac{1}{s_{\text{ICP}}} - 1$ if $s_{\text{ICP}} < 1$ and $\text{RS} = 1 - \frac{1}{s_{\text{ICP}}}$ if $s_{\text{ICP}} \geq 1$; RS = 0 indicates correct physical size.

Since the in-the-wild benchmark lacks 3D ground truths, we use the following reference-free metrics. (i) *Physical plausibility*: We calculate the hand–object closest distance (H-O Dist.) and penetration volume ratio (Pen. Ratio) to evaluate the amount of floating and intersection artifacts. (ii) *Video alignment*: We compute the Mask Intersection over Union (mIoU) between the rendered 3D silhouettes and tracked 2D masks to evaluate the spatial fidelity. (iii) *Temporal smoothness*: We report acceleration in cm/frame^2. Acc_h is the average second-order finite difference of 3D hand joints across frames. Acc_o is computed from object centers, allowing evaluation of baselines whose reconstructed object geometry varies across frames.

Implementation details. We implement our method using PyTorch [Paszke et al. 2019] and adopt the Adam optimizer [Kingma and Ba 2015] for both training the feed-forward model (Stage 2) and optimizing the hand/object poses in Stage 1 and Stage 3. Please refer to the supplementary material.

Quantitative comparison. Table 1 reports results on HO3D for both ground-truth reconstruction metrics and reference-free metrics. CHOIR achieves the best object CD, F-scores, and scale error,

Table 2. Quantitative comparison on the in-the-wild benchmark. HOLD-valid and MagicHOI-valid denote subsets where the corresponding baseline reconstructs at least one third of frames. Numbers in parentheses indicate subset sizes.

Method	Evaluation set	mIoU [%] ↑			Pen. Ratio [%] ↓			H-O [cm] ↓			Acc_h [cm/fr^2] ↓			Acc_o [cm/fr^2] ↓		
		Mean	Std	Med.	Mean	Std	Med.	Mean	Std	Med.	Mean	Std	Med.	Mean	Std	Med.
HOLD	HOLD-valid (30)	27.19	29.08	10.18	1.08	1.81	0.06	194.71	528.52	5.62	36.03	114.39	4.15	32.61	73.61	3.11
Ours	HOLD-valid (30)	78.85	11.40	83.43	6.12	5.28	4.26	0.09	0.08	0.06	0.43	0.25	0.41	0.12	0.08	0.10
MagicHOI	MagicHOI-valid (22)	23.20	21.88	16.59	3.95	6.41	1.07	6.82	10.53	2.20	3.71	5.15	2.02	2.70	5.10	1.18
Ours	MagicHOI-valid (22)	77.89	11.62	80.56	6.51	5.99	4.26	0.11	0.08	0.06	0.36	0.21	0.36	0.11	0.08	0.10
Ours	**All videos (100)**	**76.65**	**11.93**	**79.13**	**5.81**	**4.92**	**4.00**	**0.10**	**0.13**	**0.06**	**0.42**	**0.26**	**0.41**	**0.12**	**0.08**	**0.10**

indicating more accurate object shape and 6D pose recovery. MagicHOI [Wang et al. 2025] obtains a lower MPJPE by freezing the HaMeR [Pavlakos et al. 2024] hand estimate, while our method uses HaMeR only as initialization and jointly optimizes hand and object alignment. The lower penetration ratio, smaller hand-object distance, and smoother hand motion further show the benefit of contact-aware optimization beyond object reconstruction.

Table 2 reports mean, standard deviation, and median for in-the-wild videos with no 3D ground truths. HOLD and MagicHOI produce usable reconstructions on 30 and 22 videos, respectively; most other sequences fail due to COLMAP-based tracking breakdowns. For a fair comparison, we evaluate CHOIR on the same baseline-valid subsets and on the full benchmark. Even on their valid subsets, the baselines show low median mIoU and large hand-object distances, while CHOIR maintains stronger alignment and smoother hand/object trajectories for both subset and full evaluations.

Qualitative comparison. We first compare CHOIR with baseline methods on HO3D and in-the-wild videos. Figure 10 shows HO3D comparisons, and Figure 5 compares CHOIR with HOLD and MagicHOI on TASTE-Rob and self-captured videos, where the baseline pipelines produce usable reconstructions. Compared with baselines that often suffer from missing contact, object interpenetration, or floating hands, our contact-aware optimization produces tighter hand–object alignment and fewer physically implausible penetration artifacts. Figure 9 further highlights challenging in-the-wild cases handled by our method, including small objects in the 1st column, changing hand poses and contacts over time in the 2nd column, and transparent objects with small contact areas in the 3rd column. The visualized contact maps reveal the plausible hand-object contact regions localized by CHOIR. The supplementary webpage provides multi-view video comparisons against HOLD and MagicHOI, together with temporal visualizations of the contact maps. These visualizations show that CHOIR maintains coherent hand-object geometry and contact over time. EasyHOI is included quantitatively but omitted from videos due to unstable per-frame outputs.

Open-world/In-the-wild discussion. We clarify the scope of "open-world" and "in-the-wild" in four aspects. (i) For *objects*, CHOIR does not require a category-specific template or a known 3D shape; the object geometry is reconstructed from the video, thus allowing unseen everyday objects. (ii) For *camera*, CHOIR operates on monocular RGB videos without requiring a calibrated static setup, and reconstructs hand-object motion in a consistent camera/reconstruction frame. (iii) For *object motion*, we recover a rigid 6D pose trajectory, including both rotation and translation, instead of only tracking

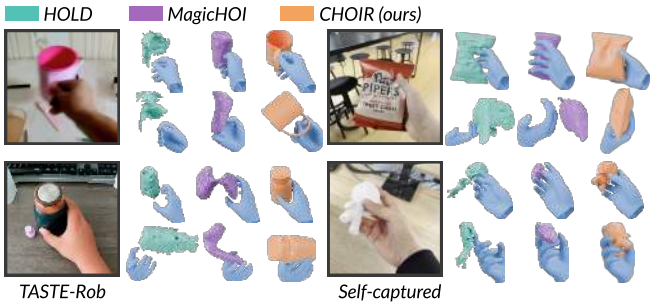

Fig. 5. Comparison with HOLD and MagicHOI on in-the-wild videos.

image-space masks or object centers. (iv) For *contact*, our formulation builds contact evidence on contact-bearing frames and updates soft contact constraints during optimization, so the active hand vertices can vary during the interaction. However, we do not handle arbitrary re-grasping or highly non-rigid object deformations.

Ablation studies. We conduct ablation studies on the in-the-wild benchmark in Table 3, where reference-free metrics directly measure visual alignment, hand-object proximity, penetration, and temporal smoothness. Rows (a) and (b) show that the Stage 1 terms $\mathcal{L}_{\text{rep}}$ and $\mathcal{L}_{\text{attr}}$ are important for isolated object fitting: removing them substantially reduces mIoU and increases hand-object distance. Row (c) removes Stage 2 rectification; although the silhouette score remains similar, the hand-object distance roughly doubles, confirming that this stage mainly corrects relative placement. Rows (d)–(f) isolate the Stage 3 penetration, contact, and temporal terms, showing their effects on penetration handling, contact alignment, and motion smoothness. We interpret penetration ratio together with hand-object distance, since a low penetration score can also come from overly separated hand and object reconstructions. Figure 8 shows that removing key components leads to inferior object poses, missing contacts, or larger penetration artifacts. Figure 7 further shows that Stage 2 rectification improves the hand-object alignment before contact correspondences are constructed. The HO3D ablation is provided in the supplementary material for completeness.

6 Conclusion

We presented CHOIR, an end-to-end framework for *open-world* 4D hand–object interaction reconstruction from monocular videos. Our central enabler is a *contact-first* formulation: we model contact as the transferable quantity to couple hand articulation and object geometry, allowing robust joint recovery even when initial monocular estimates are noisy and occluded. Combining generative HOI spatial rectification with contact-aware test-time optimization, CHOIR

Table 3. Ablation study on the in-the-wild benchmark.

Method	mIoU	Pen. Ratio	H-O Dist.	Acc_h	Acc_o
	[%] ↑	[%] ↓	[cm] ↓	[cm/fr^2] ↓	[cm/fr^2] ↓
(a) Stage 1 w/o $\mathcal{L}_{\text{rep}}$	59.24	7.82	0.13	0.43	0.09
(b) Stage 1 w/o $\mathcal{L}_{\text{attr}}$	53.06	2.50	0.42	0.45	0.08
(c) w/o Stage 2 rect.	76.78	5.30	0.22	0.36	0.12
(d) Stage 3 w/o $\mathcal{L}_{\text{pen}}$	76.65	6.37	0.10	0.42	0.12
(e) Stage 3 w/o $\mathcal{L}_{\text{contact}}$	76.65	4.61	0.12	0.42	0.12
(f) Stage 3 w/o $\mathcal{L}_{\text{temp}}$	76.65	5.90	0.11	0.43	0.12
(g) **Full (Ours)**	**76.65**	**5.81**	**0.10**	**0.42**	**0.12**

produces temporally-consistent 3D hand motion, object shape, and 6D pose trajectories, as well as explicit grasp/release events, and generalizes to unseen objects and cluttered real-world scenes.

Limitations and future work. We rely on upstream 2D and 3D initialization; errors in object segmentation/tracking or initial reconstruction can propagate to the final HOI. The optimization also depends on valid anchor frames, and degrades when objects are heavily occluded at the beginning or end of sequences. We handle only single-hand interactions and do not model bimanual manipulation or more complex dynamics. In future work, we will refine object geometry using multi-view cues from the input video, extend the framework to bimanual interactions, and move from the current iterative method toward a feed-forward model.

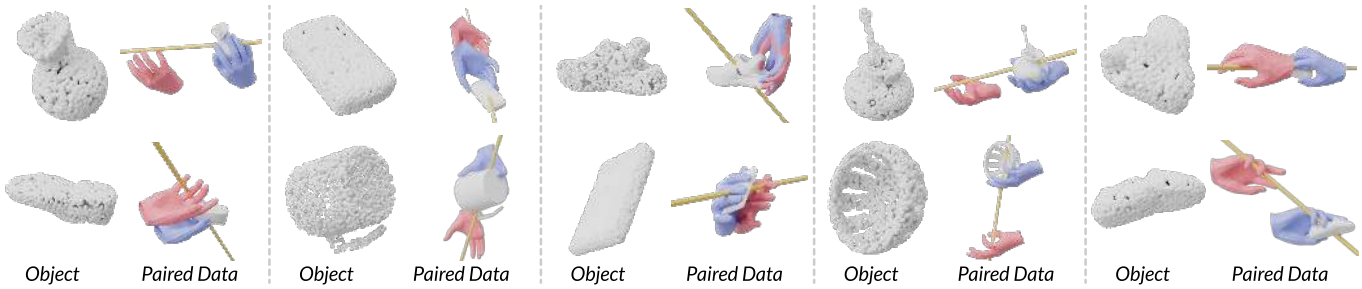

Fig. 6. Examples from our GraspPair dataset. Each case comprises a posed object point cloud with 4,096 points, a source hand pose (red), a target hand pose (blue), and the camera ray direction (yellow).

Fig. 7. Visualization of Stage 2 rectification and contact correspondences on HO3D, TASTE-Rob, and self-captured videos. We compare the reconstruction before and after Stage 2, together with ground truths when available or the final optimized result otherwise.

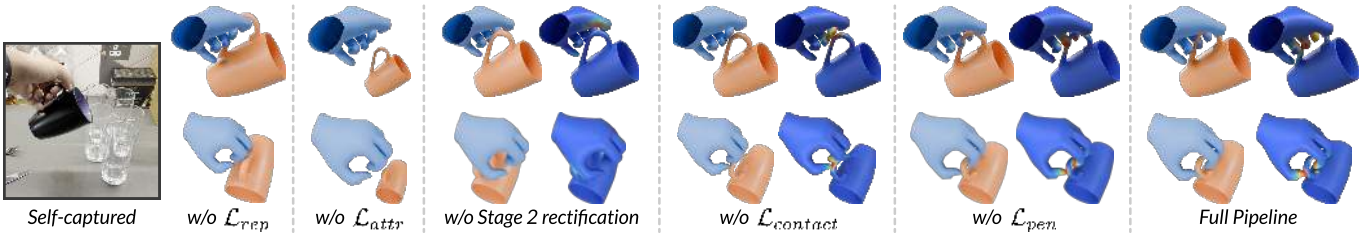

Fig. 8. Qualitative ablation of core components in CHOIR against the full pipeline. The first row shows the camera view and the second row shows a side view.

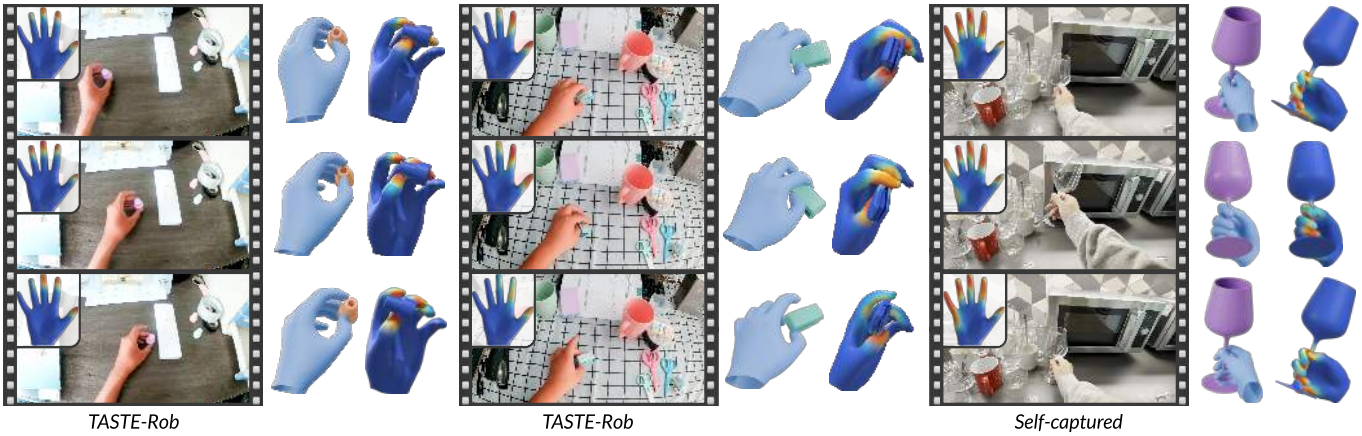

Fig. 9. Visual results of our CHOIR on the challenging cases in TASTE-Rob and self-captured videos. Each case shows the input view, a novel-view rendering with the estimated contact map, and the rest-pose hand contact map. Temporal results are shown in videos on the supplementary webpage.

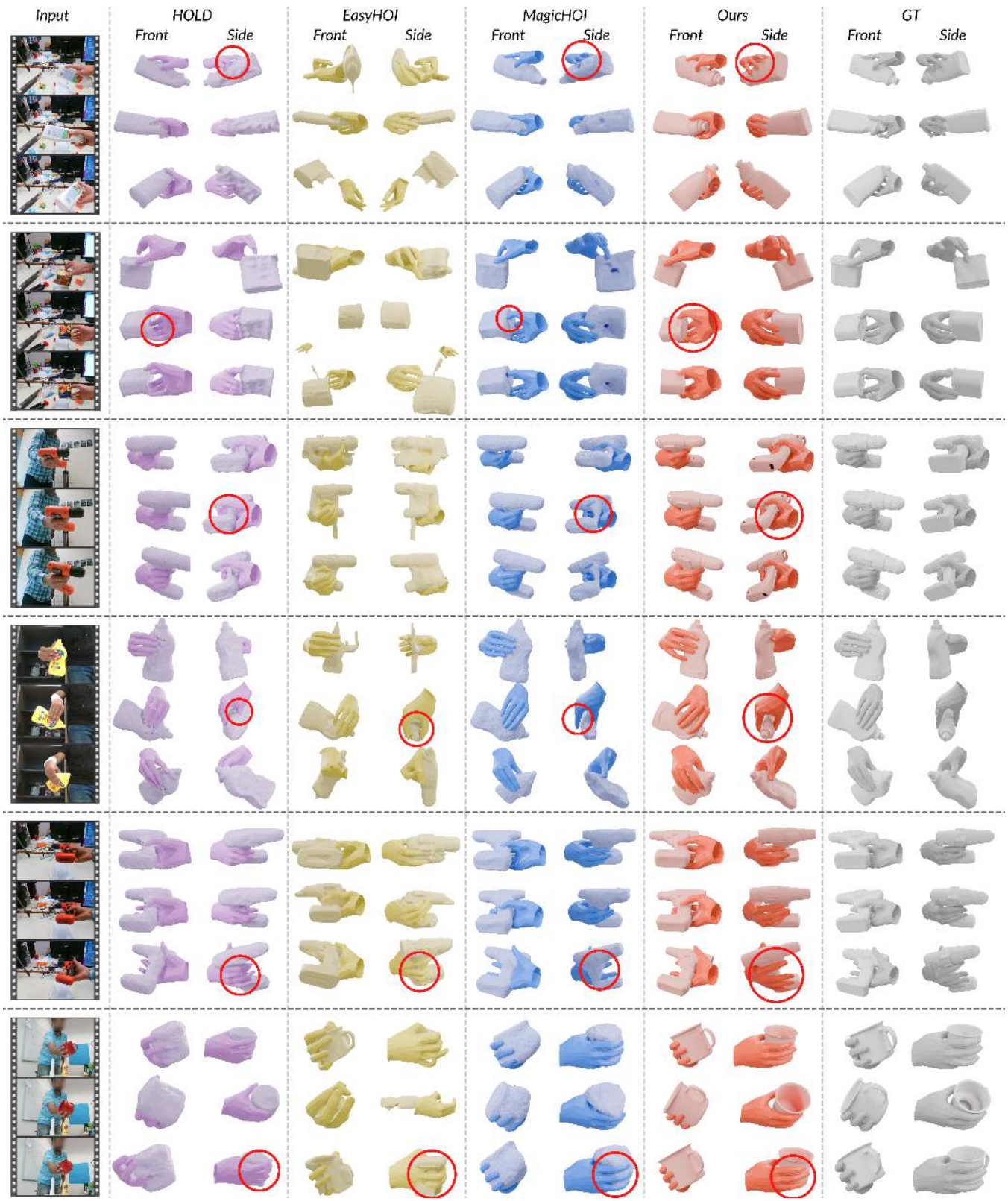

Fig. 10. Qualitative comparison between CHOIR and state-of-the-art methods on HO3D. Red circles indicate regions of interest for comparing baseline reconstructions with ours. See videos on the supplementary webpage for temporal comparisons.

Supplementary Material

7 Implementation Details

This supplementary material provides implementation details for the three stages of CHOIR, focusing on thresholds, optimization terms, schedules, and engineering choices that are omitted from the main paper due to page limit.

7.1 Stage 1: Open-World HOI Analysis

2D cue extraction. Given an RGB video, we estimate per-frame hand boxes, hand chirality, 2D hand joints, interacting-object boxes, and hand/object masks. The object pseudo-labels used for detector adaptation are mined around a motion-selected interaction frame. We identify this frame as the first local minimum in the fingertip velocity profile, which typically corresponds to the transition into grasping or manipulation. At this frame, we prompt SAM 2 [Ravi et al. 2024] with a fingertip-centered point grid (10×10) and collect candidate object masks. We keep candidates whose temporal motion is correlated with the hand trajectory, producing object pseudo-labels tied to the manipulated object rather than background distractors. These pseudo-labels are then used to fine-tune WiLoR [Potamias et al. 2025] with an interacting-object box head.

Mask propagation and amodal refinement. At test time, we initialize the object mask from the frame with the highest object-detection confidence and propagate it bidirectionally through the video using SAM 2 [Ravi et al. 2024]. Because propagated masks are modal and may miss hand-occluded regions, we refine them with an amodal video segmenter [Chen et al. 2025]. The resulting amodal silhouettes are used by the object pose tracker and by the Stage 1 and Stage 3 silhouette losses.

Hand and object initialization. HaMeR [Pavlakos et al. 2024] initializes MANO [Romero et al. 2017] hand parameters in each frame. VIPE [Huang et al. 2025] estimates camera extrinsics, and DynHaMR [Yu et al. 2025] temporally stabilizes the hand trajectory. For the object, SAM-3D-Objects [SAM 3D Team et al. 2025] reconstructs a canonical mesh and an initial metric-scale 6D pose from an object anchor frame, which defaults to the first frame but may be replaced by any frame with reliable object visibility. This anchor-frame reconstruction defines the fixed object geometry used by the follow tracker.

Guarded SAM-3D-Objects follow tracking. SAM-3D-Objects [SAM 3D Team et al. 2025] is a single-frame reconstructor. We therefore use it in a guarded follow-tracking procedure to obtain a dense object pose sequence. First, on the object anchor frame, we run full single-frame reconstruction and store the inferred object state. For subsequent frames, we initialize from this anchor-frame state, keep the shape, scale, and translation-scale latents fixed, and update only the rotation and translation pose latents. This preserves object geometry and metric scale while allowing each frame to re-estimate pose from visible image evidence. Optionally, we apply a weak pose-key velocity regularizer during follow inference:

$$\mathbf{v}_t^{(R^o,T^o)} \leftarrow \mathbf{v}_t^{(R^o,T^o)} - \lambda_{\text{temp}} \left(\mathbf{x}_t^{(R^o,T^o)} - \hat{\mathbf{x}}_{t,\text{prev}}^{(R^o,T^o)} \right), \tag{9}$$

Here $\mathbf{x}_t^{(R^o,T^o)}$ denotes the current rotation/translation pose latent and $\hat{\mathbf{x}}_{t,\text{prev}}^{(R^o,T^o)}$ denotes the propagated pose-key reference, with $\lambda_{\text{temp}} = 0.5$. The primary safeguard is an angular guard: a candidate frame is rejected if its quaternion angular distance from the last accepted tracking estimate exceeds 60°. We retry with incremented random seeds up to three times; if all attempts fail, the frame is left missing and the last accepted estimate is not updated. Missing rotations are filled by spherical linear interpolation (SLERP) and translations by linear interpolation. The resulting dense object poses initialize Stage 1 isolated sequence fitting (Step 4).

Static object initialization. Before full-sequence fitting, we refine a shared object pose and scale on the visible static segment when available. The objective combines a soft silhouette mask loss and a metric depth loss:

$$\mathcal{L}_{\text{static}} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{depth}}. \tag{10}$$

Here $\mathcal{L}_{\text{mask}}$ is the MSE between the rendered silhouette and the modal object mask, and $\mathcal{L}_{\text{depth}}$ is the MSE between rendered depth and the MoGeV2 metric-depth estimate used by SAM-3D-Objects, evaluated on pixels that are both rendered and inside the modal mask. The two terms are EMA-normalized and assigned equal relative weights. We optimize for 500 iterations with Adam [Kingma and Ba 2015] at learning rate 10^{-3}, followed by an ICP sanity check accepted only when the induced scale change lies in $[0.7, 1.3]$ and the amodal-mask IoU remains at least 0.8.

Isolated sequence fitting. Stage 1 isolated fitting refines the object trajectory and MANO [Romero et al. 2017] hand parameters before Stage 2 HOI spatial rectification. It uses the main-paper repulsion and attraction terms for object silhouette fitting, together with temporal/static object regularization and hand pose regularization. For the object, we optimize

$$\mathcal{L}_o^{\text{iso}} = \mathcal{L}_{\text{rep}} + \mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{temp}}^o + \mathcal{L}_{\text{stat}}^o. \tag{11}$$

We EMA-normalize the repulsion and attraction terms in implementation, and use weights 1.5, 1.0, 10, and 1 for $\mathcal{L}_{\text{rep}}$, $\mathcal{L}_{\text{attr}}$, $\mathcal{L}_{\text{temp}}^o$, and $\mathcal{L}_{\text{stat}}^o$, respectively. The object temporal term is

$$\mathcal{L}_{\text{temp}}^o = \mathcal{L}_{\text{sm-rot}} + \mathcal{L}_{\text{sm-tr}}, \tag{12}$$

where $\mathcal{L}_{\text{sm-rot}}$ and $\mathcal{L}_{\text{sm-tr}}$ impose temporal smoothness on object rotation and translation. Let $\mathbf{M}$ denote the rendered object silhouette, $\hat{\mathbf{M}}$ the target amodal object mask, and Ω the image domain. The repulsion loss penalizes rendered pixels outside the target mask:

$$\mathcal{L}_{\text{rep}} = \frac{1}{|\Omega|} \sum_{\mathbf{p}\in\Omega} \left[\max\left(\mathbf{M}_\mathbf{p} - \hat{\mathbf{M}}_\mathbf{p}, 0 \right) \right]^2. \tag{13}$$

The attraction loss draws 2,000 samples from the target-mask region and gives extra probability to target pixels not yet covered by the rendered silhouette. We sample pixels from

$$\mathcal{P}(\mathbf{p}) \propto \hat{\mathbf{M}}_\mathbf{p} + \alpha_{\text{uncover}} \hat{\mathbf{M}}_\mathbf{p} (1 - \mathbf{M}_\mathbf{p}) + \varepsilon, \tag{14}$$

where $\alpha_{\text{uncover}} = 20$. The sampled set $\mathcal{S}$ is restricted to target-mask pixels and is biased toward uncovered regions. Consistent with the main paper, we then define

$$\mathcal{L}_{\text{attr}} = \frac{1}{|\mathcal{S}|} \sum_{\mathbf{p}\in\mathcal{S}} \min_{\mathbf{v}\in\mathcal{V}} \|\mathbf{p} - \Pi(\mathbf{v})\|_2^2. \tag{15}$$

Here $\mathcal{S}$ is the sampled pixel set, $\mathcal{V}$ is the object-vertex set, and $\Pi(\cdot)$ is the camera projection. The static-consistency term ties pre- and post-interaction static segments to segment-level average object poses:

$$\mathcal{L}^o_{\text{stat}} = \sum_{S\in\{\text{pre,post}\}} \left(\|\mathbf{R}^o_S - \bar{\mathbf{R}}^o_S\|_2^2 + \|\mathbf{T}^o_S - \bar{\mathbf{T}}^o_S\|_2^2\right), \tag{16}$$

where $\mathbf{R}^o_S$ and $\mathbf{T}^o_S$ are the object rotation and translation for static segment S, and $\bar{\mathbf{R}}^o_S$ and $\bar{\mathbf{T}}^o_S$ denote their segment averages. This static consistency is assigned an inline weight of 10^2.

For the hand, we optimize

$$\mathcal{L}^{\text{iso}}_h = \mathcal{L}^h_{2D} + \mathcal{L}^h_{\text{anat}} + \mathcal{L}^h_{\text{prior}} + \mathcal{L}^h_{\text{depth}} + \mathcal{L}^h_{\text{temp}}. \tag{17}$$

Here $\mathcal{L}^h_{2D}$ is the 2D joint reprojection loss, $\mathcal{L}^h_{\text{anat}}$ is the MANO [Romero et al. 2017] twist-splay-bend anatomy constraint, $\mathcal{L}^h_{\text{prior}}$ anchors the hand to the HaMeR [Pavlakos et al. 2024] MANO pose in 6D rotation space, $\mathcal{L}^h_{\text{depth}}$ anchors wrist depth to the median metric-depth estimate inside the hand mask, and $\mathcal{L}^h_{\text{temp}}$ is first-order smoothness on 3D hand joints. We EMA-normalize $\mathcal{L}^h_{2D}$ and use unit weight for it after normalization. We use weights 5, 1, 10, and $\max(0.1\lambda^o_{\text{temp}}, 1.0)$ for $\mathcal{L}^h_{\text{anat}}$, $\mathcal{L}^h_{\text{prior}}$, $\mathcal{L}^h_{\text{depth}}$, and $\mathcal{L}^h_{\text{temp}}$, respectively. After isolated fitting, ray-scale alignment shifts the object trajectory along the camera ray so that its interaction-depth statistics better match the hand sequence, producing the coarse contact-agnostic 4D HOI sequence used by Stage 2.

7.2 Stage 2: Generative HOI Spatial Rectification

Generative HOI spatial rectification architecture. The rectification network follows the architecture illustrated in the main paper. Training pairs are derived from DexGraspNet [Wang et al. 2022] and filtered with PyBullet [Coumans and Bai 2016]; We uniformly sample object surface points $\mathbf{P}^o \in \mathbb{R}^{N\times 3}$ and normals $\mathbf{n}^o$, normalize the object to a unit sphere, and build dense geometry tokens from N' farthest-downsampled points using a lightweight point backbone inspired by PointNet++ [Qi et al. 2017]. The encoder is coordinate-sensitive rather than rotation-invariant, because contact validity depends on absolute object orientation.

At flow-matching time τ [Lipman et al. 2023], the scalar state z_τ represents the depth offset along the camera ray. We first embed the time, state, noisy hand joints, object scale, and viewing ray as $\mathbf{e}_\tau$, $\mathbf{e}_z$, $\mathbf{e}_J$, $\mathbf{e}_s$, and $\mathbf{e}_r$ using token-specific MLPs. The state and condition tokens are then formed as

$$\begin{aligned} \mathbf{H}_0 &= \text{Concat}\big(\mathbf{e}_\tau, \mathbf{e}_z, \mathbf{e}_J, \mathbf{e}_s, \mathbf{e}_r\big) + \mathbf{E}_{\text{PE}}, \\ \mathbf{H}_{\text{self}} &= \text{Attn}(\mathbf{H}_0, \mathbf{H}_0, \mathbf{H}_0). \end{aligned} \tag{18}$$

We obtain object geometry tokens using a PointNet++-style point encoder [Qi et al. 2017], denoted by $\mathcal{E}$:

$$\mathbf{H}_{\text{obj}} = \mathcal{E}(\mathbf{P}^o, \mathbf{n}^o) + \mathbf{E}^o_{\text{PE}}, \tag{19}$$

and the state/condition tokens query them through cross-attention:

$$\mathbf{H}_{\text{cross}} = \text{Attn}(\mathbf{H}_{\text{self}}, \mathbf{H}_{\text{obj}}, \mathbf{H}_{\text{obj}}). \tag{20}$$

Here $\text{Attn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \text{softmax}(\mathbf{Q}\mathbf{K}^\top/\sqrt{d})\mathbf{V}$ and $\mathbf{E}_{\text{PE}}$ denotes positional embedding, following standard Transformer [Vaswani et al. 2017]. The flow-matching time τ is also injected into Transformer blocks via AdaLN-style conditioning [Peebles and Xie 2023]. The token corresponding to z_τ is decoded into the scalar velocity $v_\theta(z_\tau, \tau \mid \mathbf{J}^h, \mathbf{r}, s, \mathbf{P}^o, \mathbf{n}^o)$.

Initial anchor construction. After applying the generative HOI ray-depth correction to each interaction frame, we construct initial contact correspondences on the rectified hand-object geometry. We align each per-frame object mesh to the first interaction-frame canonical mesh, sample 10,000 canonical object-surface points with a fixed seed, and search the $K = 50$ nearest samples for each candidate MANO [Romero et al. 2017] contact vertex. A correspondence is kept only if its distance is below 2 cm and the hand-normal/object-direction angle lies within a 60° cone. The accepted point is projected back to an object face and stored as face index plus barycentric coordinates.

7.3 Stage 3: Contact-aware HOI Optimization

Phase segmentation. Each video is divided into pre-static, approach, interacting, release, and post-static phases when present. The segmentation combines object point-trajectory motion from CoTracker [Karaev et al. 2025] with frame-to-frame IoU changes of the amodal object masks. Continuous-contact clips may contain only the interacting phase; missing phases are skipped rather than forced.

Objective expansion. We detail the five loss families used in Stage 3: contact $\mathcal{L}_{\text{con}}$, penetration $\mathcal{L}_{\text{pen}}$, silhouette $\mathcal{L}_{\text{sil}}$, anchor $\mathcal{L}_{\text{anc}}$, and temporal $\mathcal{L}_{\text{temp}}$. These terms are applied as test-time reconstruction constraints over a 4D sequence.

In implementation, the contact family contains the soft correspondence term and lightweight cache-stabilization terms:

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{template}} + \mathcal{L}_{\text{bridge}} + \mathcal{L}_{\text{gap}} + \mathcal{L}_{\text{patch}}. \tag{21}$$

$\mathcal{L}_{\text{contact}}$ is the soft correspondence loss that pulls active hand contact vertices toward their barycentric object anchors. $\mathcal{L}_{\text{template}}$, $\mathcal{L}_{\text{bridge}}$, $\mathcal{L}_{\text{gap}}$, and $\mathcal{L}_{\text{patch}}$ stabilize reliable contacts across short temporal gaps and local neighborhoods. We use weight 10^3 for $\mathcal{L}_{\text{contact}}$; the auxiliary stabilizers use annealed weights and are included in the released code.

The soft contact term is restricted to interaction frames:

$$\mathcal{L}_{\text{contact}} = \frac{\sum_{t,i} c_{t,i} \sum_{k=1}^{8} \omega_{t,i,k} \|\mathbf{v}^h_{t,i} - \mathbf{a}^o_{t,i,k}\|_2^2}{\sum_{t,i} c_{t,i}}. \tag{22}$$

Here $\mathbf{a}^o_{t,i,k}$ are barycentric object anchors, $\omega_{t,i,k}$ are softmax weights over the top-8 anchors computed from surface-sample distances with $\sigma = 0.01$, and $c_{t,i}$ is the Stage 3 contact confidence from current proximity/normal evidence and temporal memory. The soft correspondence cache is periodically rebuilt from the current optimized geometry, seeded by Stage 2 rectified contact evidence. Each rebuild uses a 5 cm distance threshold and a 60° normal-compatibility cone, with temporal memory for stable correspondences.

The penetration term is applied to all frames and is one-sided:

$$\mathcal{L}_{\text{pen}} = \frac{1}{T|\mathcal{V}^h|} \sum_t \sum_{\mathbf{v}^h \in \mathcal{V}^h_t} \min\big(\max((\mathbf{v}^o_{\text{nn}} - \mathbf{v}^h)^\top \mathbf{n}^o_{\text{nn}} - \epsilon_{\text{pen}}, 0), 0.04\big), \tag{23}$$

Table 4. Evaluation-set statistics. "Total cases" denotes the available cases, while method columns report the number of successfully evaluated cases.

Dataset	Total cases	3D GT	HOLD	MagicHOI	Ours
HO3D	14	Yes	14	14	14
TASTE-Rob	70	No	22	14	70
Self-captured	30	No	8	8	30
Total	114	–	44	36	114

where $\mathcal{V}_t^h$ is the hand-vertex set at frame t; $\mathbf{v}_{\text{nn}}^o$ and $\mathbf{n}_{\text{nn}}^o$ denote the nearest object vertex and its outward normal for the hand vertex $\mathbf{v}^h$, with dead zone $\epsilon_{\text{pen}} = 5\,\text{mm}$, per-vertex residual clipping threshold 0.04, and inline weight 500.

The anchor family is decomposed as

$$\mathcal{L}_{\text{anc}} = \mathcal{L}_{2D}^h + \mathcal{L}_{\text{anat}}^h + \mathcal{L}_{\text{pose}}^h + \mathcal{L}_{\text{pose}}^o. \tag{24}$$

We use weights 0.5, 30, and 100 for $\mathcal{L}_{2D}^h$, $\mathcal{L}_{\text{anat}}^h$, and $\mathcal{L}_{\text{pose}}^h$, respectively. The object pose-anchor term is split as

$$\mathcal{L}_{\text{pose}}^o = \mathcal{L}_{\text{obj-anchor}}^{\text{non-inter}} + \mathcal{L}_{\text{obj-anchor}}^{\text{inter}}, \tag{25}$$

with inline weight 100 on each component. The non-interaction component anchors object poses in pre-/post-static or non-contact frames to the Stage 1 isolated trajectory, while the interaction component anchors interaction-frame object poses to the Stage 2 rectified initialization. This prevents contact forces from drifting the object away from image evidence. The silhouette group $\mathcal{L}_{\text{sil}}$ aligns the rendered object silhouette with the amodal object mask and uses inline weight 500.

The temporal group is

$$\begin{aligned}\mathcal{L}_{\text{temp}} = {} & \mathcal{L}_{\text{pose-vel}} + \mathcal{L}_{\text{obj-vel}} + \mathcal{L}_{\text{wrist-anchor}} \\ & + \mathcal{L}_{\text{hand-tr-vel}} + \mathcal{L}_{\text{root-R-vel}} + \mathcal{L}_{\text{pose-acc}} \\ & + \mathcal{L}_{\text{hand-tr-acc}} + \mathcal{L}_{\text{root-R-acc}}.\end{aligned} \tag{26}$$

$\mathcal{L}_{\text{wrist-anchor}}$ anchors wrist translation to the rectified/input trajectory. Subscripts "vel" and "acc" denote first- and second-order temporal finite differences, while "R" and "tr" denote rotation and translation. The corresponding inline weights are 500, 500, 200, 200, 500, 1,000, 5,000, and 3,000 in the order written above. MANO [Romero et al. 2017] finger-pose and root-rotation smoothness are computed in 6D-rotation or rotation-matrix space to avoid axis-angle wrap-around artifacts. Stage 3 runs for 800 iterations with Adam [Kingma and Ba 2015] using learning rates $3{\times}10^{-4}$ for object pose, $5{\times}10^{-4}$ for MANO finger pose, and $5{\times}10^{-5}$ for wrist rotation/translation.

8 Evaluation Details

Evaluation set statistics. Table 4 summarizes the number of available and evaluated cases for each dataset and method. The in-the-wild benchmark contains 70 TASTE-Rob videos and 30 self-captured videos. We count a baseline sequence as successfully reconstructed when at least one third of its frames produce usable reconstructions; under this criterion, HOLD succeeds on 30 sequences and MagicHOI succeeds on 22 sequences, while CHOIR is evaluated on all 100 videos.

Reference-free metric details. For datasets without 3D ground truth, we report reference-free image alignment, physical plausibility, and temporal smoothness metrics. Let $\mathcal{T}$ be the set of frames with valid reconstructions for a method. For baselines with missing outputs, all per-frame metrics are averaged over $\mathcal{T}$, and temporal accelerations are computed only on consecutive valid frame triplets.

For video alignment, let $\mathbf{M}_t^{\text{rend}}$ be the rendered hand-object silhouette and $\mathbf{M}_t^{\text{obs}}$ be the tracked 2D hand-object mask at frame t. We compute

$$\text{mIoU} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{|\mathbf{M}_t^{\text{rend}} \cap \mathbf{M}_t^{\text{obs}}|}{|\mathbf{M}_t^{\text{rend}} \cup \mathbf{M}_t^{\text{obs}}|}. \tag{27}$$

For hand-object proximity, let $\mathcal{V}_t^h$ and $\mathcal{V}_t^o$ denote the reconstructed hand and object vertices in frame t. The hand-object distance is the closest surface distance,

$$\text{H-O Dist.} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \min_{\mathbf{v}^h \in \mathcal{V}_t^h,\, \mathbf{v}^o \in \mathcal{V}_t^o} \|\mathbf{v}^h - \mathbf{v}^o\|_2. \tag{28}$$

For penetration, let $d_t^o(\mathbf{v}^h)$ be the signed distance from a hand vertex to the object, with positive values inside the object and non-positive values outside. We report the percentage of hand vertices inside the object,

$$\text{Pen. Ratio} = \frac{100}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{|\mathcal{V}_t^h|} \sum_{\mathbf{v}^h \in \mathcal{V}_t^h} \mathbb{1}\left[d_t^o(\mathbf{v}^h) > 0\right]. \tag{29}$$

Finally, for temporal smoothness, let $\mathbf{J}_{t,j}^h$ be the 3D position of hand joint j at frame t, with J denoting the number of hand joints. We define the object center $\mathbf{c}_t^o$ as the centroid of the reconstructed object vertices $\mathcal{V}_t^o$, and let $\mathcal{A} \subset \mathcal{T}$ be the set of valid middle frames whose neighbors $t-1$ and $t+1$ are also valid. With positions measured in centimeters, we compute

$$\text{Acc}_h = \frac{1}{|\mathcal{A}|J} \sum_{t \in \mathcal{A}} \sum_{j=1}^{J} \left\| \mathbf{J}_{t+1,j}^h - 2\mathbf{J}_{t,j}^h + \mathbf{J}_{t-1,j}^h \right\|_2, \tag{30}$$

and use object centers as a common trajectory representation for object acceleration,

$$\text{Acc}_o = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \left\| \mathbf{c}_{t+1}^o - 2\mathbf{c}_t^o + \mathbf{c}_{t-1}^o \right\|_2. \tag{31}$$

Both acceleration metrics are reported in cm/frame^2.

9 Additional Results

This section collects supplementary visual diagnostics, ablations, runtime details, and failure-case placeholders. The visual diagnostics are included here because they expose the intermediate image evidence used by the later 4D reconstruction and contact-aware optimization stages.

Stage 1 2D assets. Stage 1 produces modal masks, amodal masks, 2D hand/object boxes, and 2D hand joints as image-space evidence for the later 4D reconstruction stages. Figure 12 shows representative outputs.

HO3D ablation. In Table 5, we include the HO3D object-centric ablation for completeness. Several Stage 3 variants produce close object-centric scores because these metrics are less sensitive to contact-side finger updates than to object geometry, object pose, and hand-root/object placement.

Table 5. Supplementary HO3D ablation study.

Method	CD	F5	F10	MPJPE	CD_h	RS
	[cm] ↓	[%] ↑	[%] ↑	[mm] ↓	[cm] ↓	↓
(a) Stage 1 w/o $\mathcal{L}_{\text{rep}}$	0.771	72.41	96.62	5.84	7.24	0.38
(b) Stage 1 w/o $\mathcal{L}_{\text{attr}}$	0.833	69.88	94.79	5.76	5.65	0.25
(c) w/o Stage 2 rect.	0.770	72.15	97.27	6.02	4.49	0.10
(d) Stage 3 w/o $\mathcal{L}_{\text{pen}}$	0.770	73.26	95.96	5.55	2.41	0.10
(e) Stage 3 w/o $\mathcal{L}_{\text{contact}}$	0.771	72.57	96.62	5.55	2.41	0.10
(f) Stage 3 w/o $\mathcal{L}_{\text{temp}}$	0.768	72.75	96.90	5.55	2.41	0.10
(g) **Full (Ours)**	**0.772**	**72.33**	**96.03**	**5.55**	**2.41**	**0.10**

Bimanual extension cases. Although the main pipeline is designed for single-hand interactions, it can be adapted with simple modifications to bimanual interactions where two hands manipulate different objects or the same object, as visualized on the supplementary webpage. When the two hands interact with different objects, we run the single-hand pipeline independently for each hand-object pair. In such cases, the hand scale and 2D joint constraints help reduce monocular depth ambiguity, so the two recovered interactions often remain visually compatible; however, there is no explicit cross-hand depth constraint or joint optimization, so global consistency is not strictly guaranteed.

For the case where two hands interact with the same object, we share the object geometry reconstructed from the same anchor frame. We then run Stages 2–3 separately for the left and right hand, obtaining two optimized hand-object sequences that use the same object shape. To merge them, we compute the rigid transformation between the two optimized object poses and use this relative transform to map one hand's global pose into the coordinate frame of the other hand-object reconstruction. This produces a common-coordinate visualization of two hands interacting with the same object. This procedure is a post-hoc composition for visualization rather than a full bimanual optimization, and future work should model both hands and their contacts jointly.

Runtime. All experiments are implemented in PyTorch [Paszke et al. 2019] and run on NVIDIA RTX 4090 GPUs. We report runtime per video, averaged over clips with approximately 200 frames. The total inference time is around 10 hours for HOLD [Fan et al. 2024], around 9 hours for EasyHOI [Liu et al. 2025], around 2 hours for MagicHOI [Wang et al. 2025], and around 30 minutes for CHOIR. For our pipeline, Stage 1 takes around 20 minutes for 2D cue extraction and mask processing, around 5 minutes for object keyframe reconstruction and guarded follow tracking, and around 30 seconds for isolated sequence fitting. Stage 2 generative HOI spatial rectification takes around 10 seconds, and Stage 3 contact-aware optimization takes around 2 minutes. The one-time training cost of the generative HOI spatial rectification module is not included in per-video inference runtime.

10 Limitations

CHOIR assumes a single hand manipulating a rigid object in a monocular RGB video. It can handle clutter, occlusion, and unknown object categories, but may fail when object geometry cannot be recovered from the anchor frame, when the object is fully occluded for a long period, or when strong rotational symmetry makes the spin angle unobservable from silhouettes and contacts alone. Non-rigid objects, articulated objects, and full bimanual joint optimization are promising directions for future work.

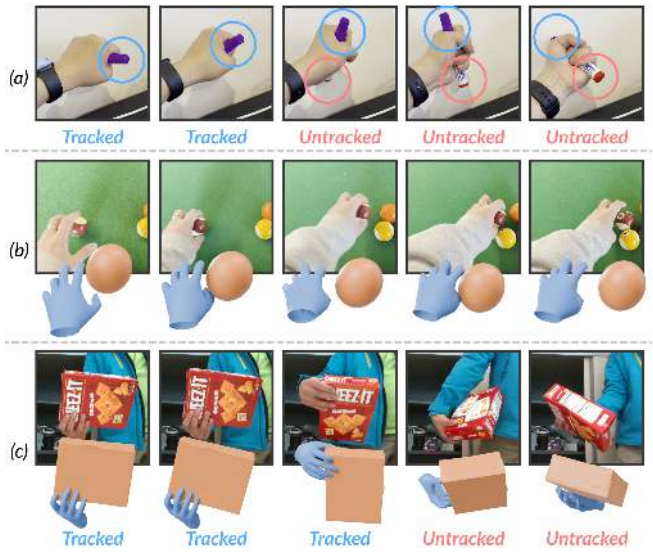

Fig. 11. Representative failure cases. Most failures originate from Stage 1 initialization: (a) incomplete 2D object detection or segmentation, (b) incorrect metric object scale estimation or scale refinement, and (c) ambiguous or inaccurate pose tracking for geometrically symmetric objects under rapid or large motion changes.

Failure cases. The main failure modes come from upstream object initialization in Stage 1. Figure 11 shows three representative categories: incomplete 2D object detection or segmentation in Step 1, incorrect metric object scale estimation or scale refinement in Step 2, and pose-tracking ambiguity for geometrically symmetric objects under rapid or large motion changes in Step 3. These errors are difficult for later stages to fully recover from because Stage 2 rectifies hand-object placement on top of the recovered object geometry, while Stage 3 optimizes around the resulting initialization and contact evidence. Incomplete masks can lead to missing or distorted object geometry, scale errors shift the interaction into an incorrect metric range, and symmetric-pose ambiguities under fast motion can make the tracked object trajectory inconsistent with the observed contact. As a result, later contact-aware optimization may improve local alignment but can still preserve an incorrect object shape, scale, or pose mode.

References

Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 2019. 3D Hand Shape and Pose from Images in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10843–10852.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. 2021. DexYCB: A Benchmark for Capturing Hand Grasping of Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaihua Chen, Deva Ramanan, and Tarasha Khurana. 2025. Using Diffusion Priors for Video Amodal Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22890–22900.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. 2022. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 231–248.

Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. 2023. Towards a Richer 2D Understanding of Hands at Scale. In *Advances in Neural*

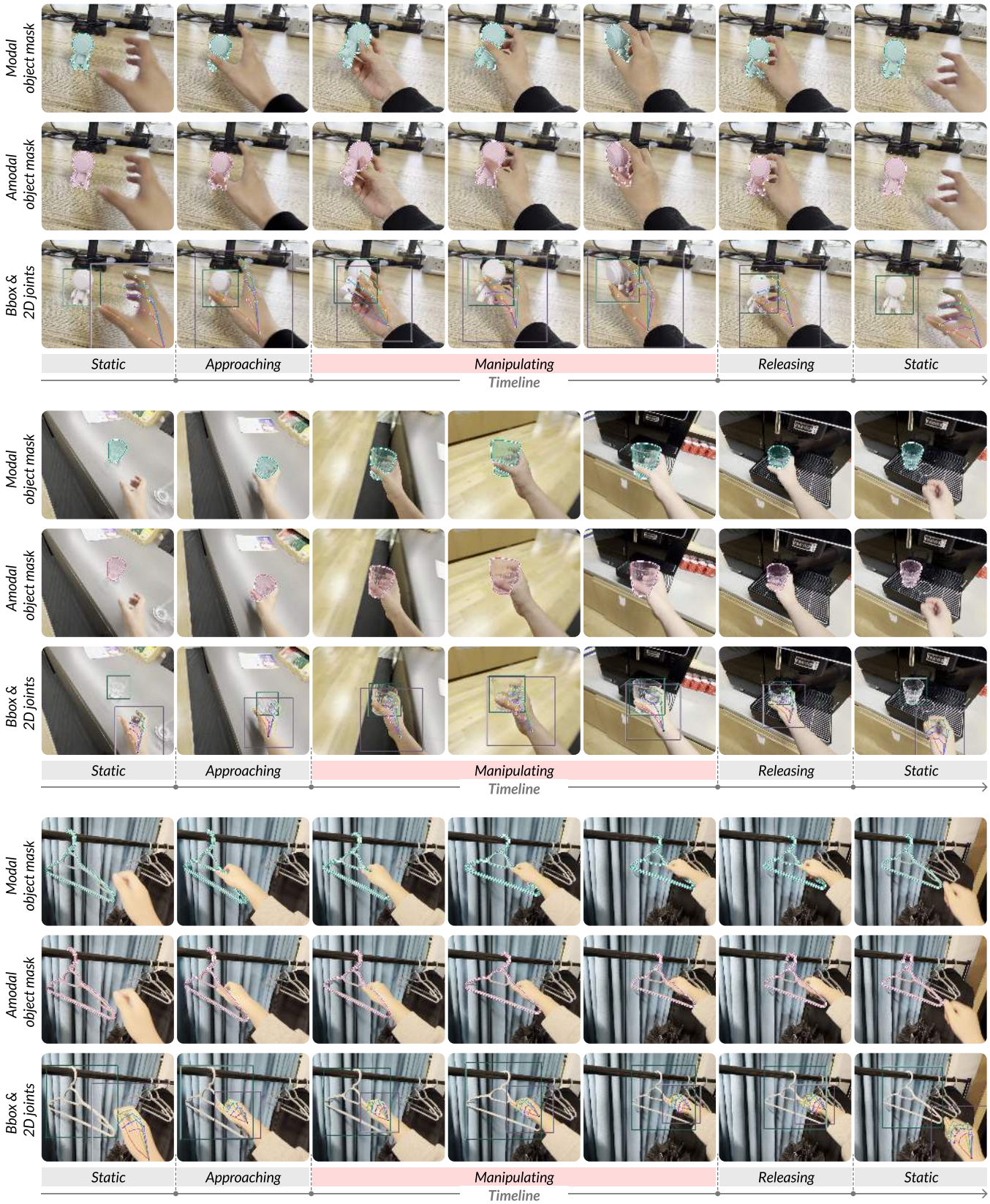

Fig. 12. Visualization of Stage 1 2D assets. The figure shows modal object masks, amodal object masks, 2D hand and object bounding boxes, and 2D hand joints, which provide the image-space evidence used by the subsequent 3D reconstruction, HOI spatial rectification, and optimization stages.

Information Processing Systems (NeurIPS), Vol. 36. 30453–30465.
Erwin Coumans and Yunfei Bai. 2016. PyBullet: A Python Module for Physics Simulation for Games, Robotics and Machine Learning. https://pybullet.org.
Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
Keyu Du, Jingyu Hu, Haipeng Li, Hao Xu, Haibin Huang, Chi-Wing Fu, and Shuaicheng Liu. 2025. Hierarchical Neural Semantic Representation for 3D Semantic Correspondence. In *Proceedings of SIGGRAPH Asia*. 1–11.
Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, and Otmar Hilliges. 2024. 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 (CVPR)*. 494–504.
Hao Feng, Zhi Zuo, Jia-hui Pan, Ka-hei Hui, Yihua Shao, Qi Dou, Wei Xie, and Zhengzhe Liu. 2025. Wonderverse: Extendable 3D Scene Generation with Video Generative Models. *arXiv preprint arXiv:2503.09160* (2025).
Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, and Kris M. Kitani. 2023. Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 23600–23611.
Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. 2021. ContactOpt: Optimizing Contact to Improve Grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1471–1481.
Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2020. HOnnotate: A Method for 3D Annotation of Hand and Object Poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. 2019. Learning Joint Reconstruction of Hands and Manipulated Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11807–11816.
Jingyu Hu, Ka-Hei Hui, Zhengzhe Liu, Hao Zhang, and Chi-Wing Fu. 2023a. CLIPXPlore: Coupled CLIP and Shape Spaces for 3D Shape Exploration. In *Proceedings of SIGGRAPH Asia*. 1–12.
Jingyu Hu, Ka-Hei Hui, Zhengzhe Liu, Ruihui Li, and Chi-Wing Fu. 2023b. Neural Wavelet-Domain Diffusion for 3D Shape Generation, Inversion, and Manipulation. *ACM Transactions on Graphics (TOG)* 42, 6 (2023).
Jingyu Hu, Ka-Hei Hui, Zhengzhe Liu, Hao Zhang, and Chi-Wing Fu. 2024. CNS-Edit: 3D Shape Editing via Coupled Neural Shape Optimization. In *Proceedings of SIGGRAPH*. 1–12.
Jingyu Hu, Bin Hu, Ka-Hei Hui, Haipeng Li, Zhengzhe Liu, Daniel Cohen-Or, and Chi-Wing Fu. 2026. PEGAsus: 3D Personalization of Geometry and Appearance. *arXiv preprint arXiv:2602.08198* (2026).
Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. 2022a. Neural Template: Topology-Aware Reconstruction and Disentangled Generation of 3D Meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18572–18582.
Ka-Hei Hui, Ruihui Li, Jingyu Hu, and Chi-Wing Fu. 2022b. Neural Wavelet-Domain Diffusion for 3D Shape Generation. In *Proceedings of SIGGRAPH Asia*. 1–9.
Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. 2025. ViPE: Video Pose Engine for 3D Geometric Perception. NVIDIA Research White Paper. arXiv:2508.10934 [cs.CV]
Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2025. CoTracker3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6013–6022.
Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.
Rosario Leonardi, Antonino Furnari, Francesco Ragusa, and Giovanni Maria Farinella. 2024. Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 36–54.
Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. In *International Conference on Learning Representations (ICLR)*.
Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. 2021. Semi-Supervised 3D Hand-Object Poses Estimation with Interactions in Time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 14687–14697.
Zhengzhe Liu, Jingyu Hu, Ka-Hei Hui, Xiaojuan Qi, Daniel Cohen-Or, and Chi-Wing Fu. 2023. EXIM: A Hybrid Explicit-Implicit Representation for Text-Guided 3D Shape Generation. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–12.
Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. 2025. EasyHOI: Unleashing the Power of Large Models for Reconstructing Hand-Object Interactions in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7037–7047.
Zishan Liu, Zecong Tang, RuoCheng Wu, Xinzhe Zheng, Jingyu Hu, Ka-Hei Hui, Haoran Xie, Bo Dai, and Zhengzhe Liu. 2026. Imagine a City: CityGenAgent for Procedural 3D City Generation. *arXiv preprint arXiv:2602.05362* (2026).
Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. MediaPipe: A Framework for Building Perception Pipelines. *arXiv preprint arXiv:1906.08172* (2019).
Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 8024–8035.
Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. 2024. Reconstructing Hands in 3D with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9826–9836.
William Peebles and Saining Xie. 2023. Scalable Diffusion Models with Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4195–4205.
Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 2025. WiLoR: End-to-End 3D Hand Localization and Reconstruction in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12242–12254.
Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. 2017. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30.
Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714* (2024).
Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics* 36, 6 (2017), 245:1–245:17.
SAM 3D Team, Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jiawei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. 2025. SAM 3D: 3Dfy Anything in Images. arXiv:2511.16624 [cs.CV] https://arxiv.org/abs/2511.16624
Dandan Shan, Jiaqi Geng, Michelle Shu, and David F. Fouhey. 2020. Understanding Human Hands in Contact at Internet Scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9869–9878.
I-Chao Shen, Ariel Shamir, and Takeo Igarashi. 2025. LayoutRectifier: An Optimization-Based Post-Processing Method for Graphic Design Layout Generation. In *Computer Graphics Forum*, Vol. 44. Wiley, e70273.
Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*. https://grab.is.tue.mpg.de
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30.
Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. 2022. DexGraspNet: A Large-Scale Robotic Dexterous Grasp Dataset for General Objects Based on Simulation. *arXiv preprint arXiv:2210.02697* (2022).
Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. 2025. MagicHOI: Leveraging 3D Priors for Accurate Hand-Object Reconstruction from Short Monocular Video Clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 5957–5968.
Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. 2020. SeqHAND: RGB-Sequence-Based 3D Hand Pose and Shape Estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 122–139.
Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. 2022. OakInk: A Large-Scale Knowledge Repository for Understanding Hand-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
Weilong Yan, Haipeng Li, Hao Xu, Nianjin Ye, Yihao Ai, Shuaicheng Liu, and Jingyu Hu. 2026. LaS-Comp: Zero-Shot 3D Completion with Latent-Spatial Consistency. *arXiv preprint arXiv:2602.18735* (2026).

Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J. Black. 2026. Predicting 4D Hand Trajectory from Monocular Videos. In *International Conference on 3D Vision (3DV)*.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. 2024. G-HOP: Generative Hand-Object Prior for Interaction Reconstruction and Grasp Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1911–1920.

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. 2022. 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 (CVPR)*. 3895–3905.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. 2023. Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 19717–19728.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. 2025. Dyn-HaMR: Recovering 4D Interacting Hand Motion from a Dynamic Camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27716–27726.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. 2024. 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 (CVPR)*. 445–456.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. 2025b. HaWoR: World-Space Hand Motion Reconstruction from Egocentric Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1805–1815.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. 2025a. BimArt: A Unified Approach for the Synthesis of 3D Bimanual Interaction with Articulated Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27694–27705.

Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. 2025. TASTE-Rob: Advancing Video Generation of Task-Oriented Hand-Object Interaction for Generalizable Robotic Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27683–27693.

Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Shi. 2020. Monocular Real-Time Hand Shape and Motion Capture Using Multi-Modal Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5346–5355.

Runsong Zhu, Yuan Liu, Zhen Dong, Yuan Wang, Tengping Jiang, Wenping Wang, and Bisheng Yang. 2021. AdaFit: Rethinking Learning-Based Normal Estimation on Point Clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 6118–6127.

Runsong Zhu, Shi Qiu, Qianyi Wu, Ka-Hei Hui, Pheng-Ann Heng, and Chi-Wing Fu. 2024a. PCF-Lift: Panoptic Lifting by Probabilistic Contrastive Fusion. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 92–108.

Runsong Zhu, Di Kang, Ka-Hei Hui, Yue Qian, Shi Qiu, Zhen Dong, Linchao Bao, Pheng-Ann Heng, and Chi-Wing Fu. 2024b. SSP: Semi-Signed Prioritized Neural Fitting for Surface Reconstruction from Unoriented Point Clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3769–3778.

Runsong Zhu, Shi Qiu, Zhengzhe Liu, Ka-Hei Hui, Qianyi Wu, Pheng-Ann Heng, and Chi-Wing Fu. 2025a. Rethinking End-to-End 2D to 3D Scene Segmentation in Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3656–3665.

Runsong Zhu, Ka-Hei Hui, Zhengzhe Liu, Qianyi Wu, Weiliang Tang, Shi Qiu, Pheng-Ann Heng, and Chi-Wing Fu. 2025b. COS3D: Collaborative Open-Vocabulary 3D Segmentation. *arXiv preprint arXiv:2510.20238* (2025).