COLLABORATIVE HUMAN-OBJECT INTERACTION RECONSTRUCTION FROM A MONOCULAR VIDEO

Anonymous authors
Paper under double-blind review



Figure 1: **COMBO** reconstructs *collaborative human-object interaction* from a monocular video. Without object templates or ground-truth cameras, it recovers humans and the object in a shared 4D space, guided by a temporal contact graph that grounds interactions as they hold, break, and re-form.

ABSTRACT

Collaborative human-object interaction (HOI) reconstruction requires recovering how several people manipulate a shared object over time. Existing video-based HOI reconstruction methods fall short in this scenario, as they neither align multiple human-object pairs in a shared 4D space nor reason about how contacts among multiple humans and the shared object evolve over time. To address this, we present COMBO, a 4D reconstruction method for reconstructing collaborative HOIs from a monocular video. Our key insight is that, in collaborative scenarios, the humans and the object are tightly coupled through contacts that evolve over time, and these time-varying contacts provide the constraint needed for joint reconstruction. We design a two-stage approach: human-anchored initialization first establishes a shared 4D space, using human motions to calibrate depth, and initializes the object. Subsequently, temporal contact-guided optimization builds a proximity-based temporal contact graph from 2D and 3D cues, without a learned contact predictor. Unlike static or single-pair contact formulations, our graph captures temporally varying contacts across participants, guiding the joint alignment of all entities in a shared 4D space. Experiments on the CORE4D benchmark show lower reconstruction errors than competing video-based methods. Our results on in-the-wild and generated videos demonstrate the generalization capability of COMBO. *Code will be publicly available upon acceptance.*

1 INTRODUCTION

Humans routinely interact and collaborate with objects in daily life: they move a table together or help lift a heavy box, or pass an object from one person to another (Fig. 1). Reconstructing such collaborative human-object interaction (HOI) in 4D from a monocular video is fundamental to understanding how humans coordinate through objects, with applications in human-robot collaboration,

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.

humanoid manipulation, and AR/VR content creation. Recent datasets (Liu et al., 2025; Zhang et al., 2024) capture these interactions with dense multi-view cameras and body-attached sensors, but these dedicated systems are hard to scale to the diversity of everyday collaboration. Monocular videos offer a practical alternative, and recent 3D foundation models for humans, depth, and objects (Sárándi & Pons-Moll, 2024; Lin et al., 2026; Chen et al., 2026) can estimate each entity from them, but only in isolation, *i.e.*, they estimate each entity independently, ignoring the collaborative context.

Collaborative HOI reconstruction must couple the motions of all humans through a single shared object trajectory as their contacts change over time. When two humans carry a table, their contacts simultaneously constrain its motion; during a handover, these constraints shift between them as they grasp and release the object. Ignoring these changes can keep a released hand constrained to the object or leave a new contact unconstrained. Joint reconstruction therefore accounts for both simultaneous contacts and changes in which humans and hands constrain the object.

Existing methods address only parts of the requirements for collaborative HOI reconstruction. Single-person HOI methods (Xie et al., 2026), when applied independently to each human, do not explicitly enforce a single shared object trajectory consistent with all participants. Collaborative reconstruction methods (Kogashi et al., 2026) operate on individual images, while interaction generation methods (Li & Dai, 2026) use text-derived contact relations that may not match the observed interactions. These limitations motivate estimating time-varying contacts from the video and using them to jointly refine all humans and the shared object.

To mitigate this, we propose COMBO, a training-free framework that first establishes a shared 4D space without relying on contacts and then identifies the contacts within it to couple all entities (Fig. 2). The first stage, *human-anchored initialization*, establishes the shared 4D space. We refine the human motions over the sequence and use their body scale as a metric prior to calibrate depth. The object, reconstructed without a template, is then initialized in this human-anchored space. The initialization places all entities in a common coordinate frame but does not yet enforce consistency between their motions and contacts. The second stage, *temporal contact-guided optimization*, couples and refines the entities within the shared 4D space. We build a temporal contact graph whose frame-wise edges, inferred from complementary 2D and 3D proximity cues, specify which hand of which human contacts the object. The graph requires no learned contact predictor and naturally captures contacts that hold, break, and re-form across humans, over time. Guided by the graph, we first optimize the object motion with the humans fixed, then jointly refine the motion of the object and all human entities to improve the accuracy and the physical plausibility of the reconstruction.

We evaluate COMBO on a collaborative HOI benchmark, and demonstrate more accurate 4D HOI than competing methods. We investigate the effect of the key components through ablations and demonstrate its generalization capability to in-the-wild and generated videos. Our contributions are:

- We present COMBO, the first training-free framework that reconstructs collaborative HOI from a monocular video, recovering multiple humans and the object in a shared 4D space without an object template or ground-truth cameras.
- We introduce a two-stage design that first establishes the human-anchored shared 4D space and then refines the shared object trajectory and human motions guided by their contacts.
- We introduce a proximity-based temporal contact graph that specifies which hand of which person contacts the object at each frame without a learned contact predictor, and applies the corresponding contact constraints to collaborative refinement.

2 Related Work

2.1 Human-scene reconstruction

Recent 3D foundation models (Wang et al., 2024; Leroy et al., 2024; Wang et al., 2025a; 2026a;b; 2025b; Lin et al., 2026) reconstruct scene geometry and camera parameters from images or videos. Building on these models, recent efforts have extended reconstruction to humans jointly with their surrounding scenes. From multi-view images, HSfM (Müller et al., 2025) jointly recovers multiple human meshes, the scene point cloud, and the cameras, while PhySIC (Yalandur-Muralidhar et al., 2025) and GRAFT (Pradyumna et al., 2026) reconstruct humans and the scene jointly from a single image via optimization and learned refinement, respectively. For monocular videos, JOSH (Liu et al.,

2026) jointly optimizes human motion, scene geometry, and camera poses with human-scene contact constraints, while Human3R (Chen et al., 2025) and UniSH (Li et al., 2026) regress them in a single feed-forward pass. These methods place humans and scenes in a shared 3D space, but they either merge the manipulated object into the scene or discard it, so they recover neither its full shape nor its six-degree-of-freedom (6-DoF) motion. In contrast, we reconstruct the interacting object as a separate mesh in a category-agnostic manner and track its 6-DoF motion jointly with the humans.

2.2 Single human-object interaction reconstruction

Single human-object interaction reconstruction from monocular inputs has progressed with 3D HOI datasets (Bhatnagar et al., 2022; Zhang et al., 2023; Huang et al., 2024; Cseke et al., 2025). Image-based methods regress the human and the object using interaction priors (Zhang et al., 2020; Xie et al., 2022; Nam et al., 2024). Video-based methods track them with a given object template or within limited object categories (Xie et al., 2023; 2025). CARI4D (Xie et al., 2026) and RHINO (Xue et al., 2026) extend reconstruction to novel object categories, and RePHO (Huang et al., 2026) and Real2Sim (Zhao et al., 2026) leverage physics-based refinement. Several methods use contact to couple the human and the object (Xie et al., 2022; Nam et al., 2024; Xie et al., 2026), and dedicated contact estimators exist (Tripathi et al., 2023; Yang et al., 2024; Dwivedi et al., 2025). However, these estimators take a *static* image, and CARI4D predicts per-frame contact only for the hands of one person. All of these methods assume that a *single* person manipulates the object, so they cannot model collaboration, where several people hold or hand over the object. On the other hand, HOI-PAGE (Li & Dai, 2026) generates collaborative HOI motions from text prompts by reasoning about part-level contact relationships with a large language model. In contrast, we jointly reconstruct *multiple human entities and the object across time* without object templates. We encode their contacts as a temporal relation graph built from 2D and 3D proximity, and this graph guides the joint optimization.

2.3 Collaborative human-object interaction reconstruction

Collaborative multiple human-object interaction remains far less explored than single HOI, and existing work relies mainly on datasets captured with dedicated multi-view systems and pre-scanned objects (Zhang et al., 2024; Kogashi et al., 2026; Liu et al., 2025). HOI-M^3 (Zhang et al., 2024) provides large-scale recordings of multiple humans and objects in contextual environments, but does not target collaborative object manipulation. CORE4D (Liu et al., 2025) focuses on collaboration, however, its benchmarks address motion forecasting and interaction synthesis rather than reconstruction. Moreover, these methods operate on a per-frame basis, and assume known object geometries. Lacking explicit temporal reasoning and category-agnostic reconstruction of unseen objects, they cannot recover how a collaborative interaction evolves over time, *e.g.*, when an object is handed from one person to another. In contrast, we reconstruct collaborative interactions from a monocular video without privileged object template meshes, and explicitly reason about how the contacts between the humans and the object evolve over time.

3 COMBO: Collaborative Human-Object Interaction Reconstruction from a Monocular Video

We introduce COMBO, a method that reconstructs collaborative human-object interactions from a monocular video (Sec. 3.1 and Fig. 2). COMBO first establishes a human-anchored shared 4D space (Sec. 3.2) and identifies which hand of which human contacts the object at each frame (Sec. 3.3). These contacts then guide the joint refinement of all humans and the shared object (Sec. 3.4).

3.1 Overview

Our goal is to reconstruct collaborative human-object interaction (HOI) from a monocular video $\{\mathbf{I}_t\}_{t=1}^{T}$ of N humans interacting with a shared object, where $\mathbf{I}_t \in \mathbb{R}^{H\times W\times 3}$ denotes the t-th RGB frame and T is the video length. Our method, COMBO, reconstructs human motions, object geometry, and object motion in a shared 4D space. It is training-free and requires neither known object geometry nor ground-truth camera parameters. To reconstruct plausible contacts, each human and the object must be well aligned in the shared 4D space. We therefore introduce a two-stage

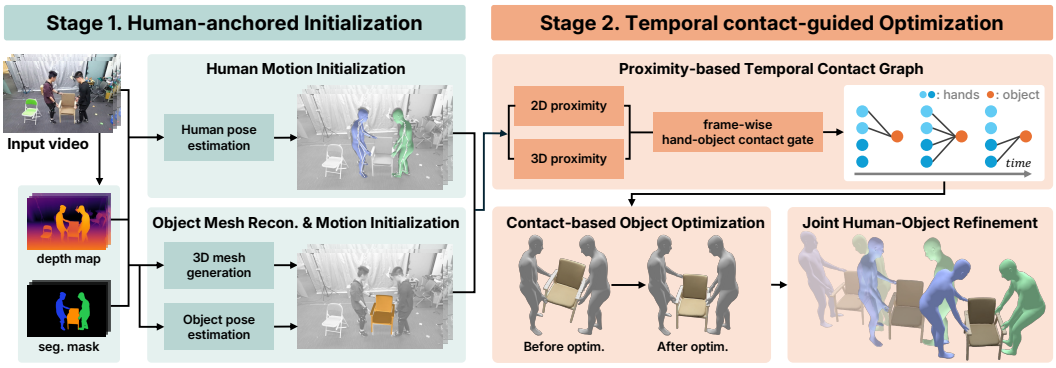

Figure 2: **COMBO pipeline.** Stage 1 recovers each entity with foundation models, initializing 4D motions of the humans and the object independently. The entities are individually plausible but not yet coupled. Stage 2 couples them through the temporal contact graph, whose hand-object edges, inferred from 2D and 3D proximity, appear and vanish as each person grasps, releases, or hands over the object. COMBO uses it to first optimize the object motion and then refine the motions of humans and objects jointly, yielding a coherent 4D reconstruction of the collaboration.

design: 1) human-anchored initialization and 2) temporal contact-guided optimization (see Fig. 2). Human-anchored initialization establishes the shared 4D space and places each human and the object independently, without using contacts. We refine the human motions over the sequence and use the human body as a spatial prior to calibrate depth, which subsequently guides object initialization. Within the shared 4D space, we build a proximity-based contact graph that specifies *temporal contact*, *i.e.*, which hand of which human contacts the object at each frame. The contact graph then drives the temporal contact-guided joint refinement of humans and objects. We first optimize the object motion with the humans fixed, then jointly refine the object motion and all humans' arm poses. In this way, time-varying contacts couple the motions of all humans through a single shared object trajectory.

3.2 Human-anchored initialization

We construct the shared 4D space for all the human and object entities by anchoring the space to humans, whose body scale serves as a metric prior. From a monocular video, we estimate per-frame depth maps and camera intrinsics K using Depth Anything 3 (Lin et al., 2026), and human and object masks using SAM3 (Carion et al., 2026). These observations support human motion refinement, depth calibration, and object initialization.

Human motion refinement. We obtain per-frame human estimates using NLF (Sárándi & Pons-Moll, 2024) and represent each human using SMPLH (Loper et al., 2015; Romero et al., 2017) parameters. A human i at frame t is represented by $\mathbf{M}_{i,t} = \{\theta_{i,t}, \phi_{i,t}, \tau_{i,t}, \beta_i\}$, with body pose $\theta_{i,t} \in \mathbb{R}^{21\times6}$, body shape $\beta_i \in \mathbb{R}^{10}$, root orientation $\phi_{i,t} \in \mathbb{R}^6$, and root translation $\tau_{i,t} \in \mathbb{R}^3$. The posed human mesh vertices are $\mathbf{V}_{i,t} \in \mathbb{R}^{6890\times3}$. To reduce frame-wise noise and tolerate momentary occlusions, we refine each human's body pose, root orientation and translation over the sequence, while keeping their estimated body shape fixed. The human refinement objective combines a confidence-weighted 2D keypoint reprojection loss and a 3D joint acceleration loss, as follows:

$$\mathcal{L}_{\text{human}} = \lambda_{\text{kpt}}\mathcal{L}_{\text{kpt}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}}, \tag{1}$$

with both terms averaged over all human entities. Please refer to Sec. A.1 for detailed loss definitions. The keypoints are obtained from Sapiens2 (Khirodkar et al., 2026) and they keep human motions aligned with the input frames, while the acceleration term suppresses abrupt temporal drift changes. The refined estimates provide a robust human initialization for the subsequent steps.

Human-anchored depth calibration. We use the temporally refined human meshes to calibrate the predicted depth to the human body scale. Specifically, we estimate one global scale $a \in \mathbb{R}$ and shift $b \in \mathbb{R}$ over the sequence by aligning the predicted depth in the human regions with the depth rendered from the human meshes. For a raw depth map $\mathbf{D}_t^{\text{raw}}$, the calibrated depth is $\mathbf{D}_t = a\mathbf{D}_t^{\text{raw}} + b$. We

apply the same calibration to the object regions, ensuring that object placement and subsequent 3D contact reasoning are conducted in a shared space.

Object initialization. We reconstruct a canonical object mesh from the first frame and its object mask using an image-to-3D (Chen et al., 2026) model. We align its scale and initial pose to the calibrated depth and object silhouette. The resulting canonical geometry $\mathbf{O}$ includes this scale and remains fixed thereafter; its posed vertices are $\mathbf{O}_t = \mathbf{R}_t\mathbf{O} + \mathbf{t}_t$, where $\mathbf{R}_t \in \mathrm{SO}(3)$ and $\mathbf{t}_t \in \mathbb{R}^3$ denote object rotation and translation.

To initialize object motion, we use the relative rotations from ProxyPose (Zhang et al., 2026), accumulate them across overlapping clips, and apply them to the aligned first-frame orientation. We initialize translation separately from the calibrated object depth at each frame, rather than from the tracker. Together, these estimates align the object trajectory in the same camera coordinate system as the humans. The initialization aligns the human and object in a shared space but does not yet enforce the physical contacts. Please refer to Sec. A.1 for alignment and motion initialization details.

3.3 Proximity-based temporal contact graph

Within the shared 4D space, we determine which hand of which human contacts the object at each frame. Humans may hold the object simultaneously, while their hands individually grasp or release the object at different times. We model these contact assignments as a temporal contact graph, capturing contact relationships that hold, break, and re-form across humans.

Temporal contact graph. Let $h_{i,s}$ denote hand $s \in \{L, R\}$ of human $i \in \{1, \dots, N\}$, and let o denote the shared object. At frame t, we define the bipartite graph $\mathcal{G}_t = (\mathcal{H}, \mathcal{O}, \mathcal{E}_t)$, where $\mathcal{H} = \{h_{i,s}\}$ denotes the hands of all humans and $\mathcal{O} = \{o\}$ the shared object. The binary gate $g_{i,s,t} \in \{0, 1\}$ indicates whether hand $h_{i,s}$ is in contact with the object at frame t. Accordingly, the active hand-object edges are defined as:

$$\mathcal{E}_t = \{(h_{i,s}, o) \mid g_{i,s,t} = 1\}. \tag{2}$$

The sequence $\{\mathcal{G}_t\}_{t=1}^{T}$ forms the temporal contact graph. Each edge specifies which human's hand is in contact with the object; the graph can represent simultaneous contacts by multiple humans and contact transfers between humans over time.

Proximity-based contact estimation. We determine each edge from geometric proximity, combining complementary 2D and 3D cues without a learned contact predictor. The 2D cue measures the proximity in the frame without relying on object depth, while the 3D cue helps distinguish actual contact from 2D overlap caused by occlusion. Let $\mathbf{V}_{i,s,t} \in \mathbb{R}^{778\times 3}$ denote the vertices of hand s of human i at frame t from the initialized human mesh, and let $\mathbf{m}_t$ denote the object mask. We obtain object-depth points $\mathbf{X}_t$ by back-projecting all valid object-mask pixels using the calibrated depth $\mathbf{D}_t$. We measure the two proximity cues as follows:

$$\begin{aligned} d^{\mathrm{2D}}_{i,s,t} &= \mathrm{dist}_{\mathrm{2D}}\big(\Pi_{K_t}(\mathbf{V}_{i,s,t}), \mathbf{m}_t\big), \\ d^{\mathrm{3D}}_{i,s,t} &= \mathrm{dist}_{\mathrm{3D}}\big(\mathbf{V}_{i,s,t}, \mathbf{X}_t\big). \end{aligned} \tag{3}$$

Here, $\Pi_{K_t}(\cdot)$ denotes perspective projection. The distance functions measure the minimum Euclidean distances from the projected hand vertices to the object-mask pixels and from the 3D hand vertices to the object-depth points, respectively.

We convert each distance to a proximity score $\phi_\sigma(d) = \exp(-d/\sigma) \in (0, 1]$, where $\sigma_{\mathrm{2D}}, \sigma_{\mathrm{3D}} \in \mathbb{R}_+$ control the decay in the respective distance units. We threshold the geometric mean of the two scores and apply a temporal median filter to obtain the binary contact gate:

$$g_{i,s,t} = \underset{\ell \in \mathcal{W}_t}{\mathrm{median}}\, \mathbb{1}\left[\sqrt{\phi_{\sigma_{\mathrm{2D}}}\big(d^{\mathrm{2D}}_{i,s,\ell}\big)\, \phi_{\sigma_{\mathrm{3D}}}\big(d^{\mathrm{3D}}_{i,s,\ell}\big)} > \gamma\right]. \tag{4}$$

Here, $\gamma \in (0, 1)$ is the contact threshold and $\mathcal{W}_t$ is the temporal filtering window at frame t. In this way, the resulting contact gate $g_{i,s,t}$ specifies the time-varying contact relationship between humans and the object, and this representation is used for the subsequent refinement step. Contact estimation is based on the initialized human hands and calibrated object depth and masks, while subsequent refinement enforces these contacts on the reconstructed human and object meshes. Please refer to Sec. A.2 for the details of the contact gate.

3.4 CONTACT-GUIDED COLLABORATIVE REFINEMENT

The proposed proximity-based temporal contact graph identifies which human hands constrain the object, but independently initialized human and object trajectories do not guarantee the contacts. We use the graph to improve the physical plausibility of their interaction while maintaining the image alignment and temporal consistency. We first refine the object with the humans fixed, then jointly adjust the object and human arms, *i.e.*, the shoulder, elbow, and wrist joints. This scheduled refinement lets the object adapt to the human motion first, and then allows both the object and the humans to move while respecting their spatial configuration based on contact estimates. Object shape geometry and scale remain fixed through the refinement steps.

Contact-based object optimization. We optimize the 6-DoF object pose $\{\mathbf{R}_t, \mathbf{t}_t\}_{t=1}^{T}$ over the sequence by minimizing the objective term:

$$\mathcal{L}_{\text{obj}} = \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{init}}\mathcal{L}_{\text{init}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{pene}}\mathcal{L}_{\text{pene}}. \tag{5}$$

The contact term collects the active edges from all human entities in the video:

$$\mathcal{L}_{\text{cont}} = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{s\in\{L,R\}} g_{i,s,t}\, d_{\text{cont}}(\mathbf{V}_{i,s,t}, \mathbf{O}_t), \tag{6}$$

where d_{cont} measures distances from the closest hand vertices to the posed object surface. At each frame, the object pose $(\mathbf{R}_t, \mathbf{t}_t)$ is optimized using contact constraints from all hands with $g_{i,s,t} = 1$. When multiple humans hold the object, their contacts jointly constrain its pose. The remaining terms keep the contact-driven correction consistent with the observations and the initialized motion. $\mathcal{L}_{\text{init}}$ combines alignment to the calibrated depth points and object masks with a prior on the initialized poses. $\mathcal{L}_{\text{temp}}$ keeps the object trajectory temporally smooth and the contacting hands static relative to the object, and $\mathcal{L}_{\text{pene}}$ penalizes human vertices inside the object.

Joint human-object refinement. Subsequent to the contact-based object optimization, we further conduct joint human-object refinement to improve the accuracy. We jointly optimize the humans and the object, since object-only optimization does not correct errors in human poses. We jointly refine the object rotation and translation and the human arm poses, *i.e.*, the shoulder, elbow, and wrist poses of all humans, by minimizing the combined objective of Eq. (1) and Eq. (5):

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{obj}} + \mathcal{L}_{\text{human}}. \tag{7}$$

The contact term penetration terms act on both the object and the human arms, while the human objective maintains keypoint alignment and temporal consistency. Human root orientation and translation, and body shape stay fixed. We denote the final human motion parameters by $\mathbf{M}^*_{i,t}$ and the final object poses by $(\mathbf{R}^*_t, \mathbf{t}^*_t)$. See Sec. A.3 for detailed loss definitions and hyperparameters.

4 EXPERIMENTS

In this section, we present qualitative and quantitative results of our method on collaborative human-object interaction benchmarks, and compare COMBO with recent competing methods. We also conduct ablation studies of our proposed pipeline components and show generalization capability with reconstruction results from in-the-wild and synthetically generated videos.

4.1 EVALUATION SETUP

We evaluate on CORE4D-Real (Liu et al., 2025), using the full test set of our benchmark for the quantitative comparisons. We report Chamfer distances for the humans (CD-h), object (CD-o), and combined human-object geometry (CD-c). We also measure human pose accuracy using PA-MPJPE and object translation and rotation errors using T_{obj} and R_{obj}, respectively. Details on dataset and evaluation metrics are in Sec. B.1 and Sec. B.3, respectively.

4.2 QUALITATIVE RESULTS

In Fig. 3, we compare our reconstructions with competing methods. HOI-PAGE (Li & Dai, 2026) shows human-object misalignment and incorrect object orientations, as in the bucket and table



Figure 3: **Qualitative comparison on CORE4D.** Visualization of our reconstruction results compared with competing methods in camera-view and side-view renderings (insets). The motions reconstruted by COMBO are well aligned with input video frames while maintaining plausible contacts and spatial configurations between multiple humans and the shared object.

Table 1: **Quantitative comparison on CORE4D.** We compare reconstruction quality in two settings: with access to the ground-truth object mesh (top) and without access to it (bottom). COMBO reconstructs human and object motions from monocular video without requiring a template mesh and achieves the best results across all metrics in the setting without the ground-truth object mesh. The best and second-best results within each setting are marked in **bold** and underlined, respectively.

Method	Input	GT object mesh	CD-h (↓)	CD-o (↓)	CD-c (↓)	PA-MPJPE (↓)	T_{obj} (↓)	R_{obj} (↓)
CARI4D	Video	✓	22.70	84.15	30.17	8.30	56.16	45.09
HOI-PAGE	Video	✓	77.85	103.53	67.79	6.85	73.95	41.94
MMHOI-Net	Image	✓	**18.11**	58.82	-	-	-	-
COMBO (Ours)	Video	✓	20.56	**28.01**	**20.03**	**6.23**	**25.58**	**17.67**
CARI4D	Video	-	23.36	134.28	37.85	8.92	84.15	47.17
HOI-PAGE	Video	-	77.85	102.46	67.69	6.85	72.49	43.77
COMBO (Ours)	Video	-	**20.50**	**28.86**	**20.16**	**6.24**	**26.39**	**17.59**

examples. CARI4D (Xie et al., 2026) produces plausible human reconstructions, but the object is not always aligned with both humans' hands, leaving visible gaps. In contrast, COMBO estimates hand-specific contacts with the proximity-based temporal contact graph and incorporates them into the joint refinement of all humans and the shared object. The resulting reconstructions more closely match the object poses in the input frames while recovering plausible contacts with both humans, yielding more coherent spatial configurations for collaborative HOI.

4.3 Quantitative Results

Table 1 reports quantitative results on CORE4D, grouped by whether the GT object mesh is available. With access to the GT object mesh (top partition), COMBO achieves the best reported results on five of the six metrics and ranks second on CD-h. More importantly, without access to the GT object mesh (bottom partition), COMBO outperforms all competing methods across all six metrics. Our template-free variant also achieves better results on five of the six metrics than the competing methods that use GT object meshes. CARI4D uses frame-wise contact predictions for a single human-object pair, without jointly constraining the shared object across multiple humans. HOI-PAGE

Table 2: **Ablation study.** Each component contributes to the reconstruction quality, and the full model with the proposed stage order achieves the lowest CD for the human, object, and combined meshes. Evaluated on a subset of 100 test sequences from CORE4D (Liu et al., 2025).

Method	CD-h (↓)	CD-o (↓)	CD-c (↓)	PA-MPJPE (↓)	T_{obj} (↓)	R_{obj} (↓)
Ours (full)	**19.23**	**25.36**	**18.85**	6.17	**24.25**	15.26
w/o human temp. refinement	19.44	25.38	18.91	6.36	24.32	**15.11**
w/o object optim. (Stage-1)	19.26	27.40	19.34	6.23	25.69	14.96
w/o joint refinement (Stage-2)	19.96	25.95	19.35	**6.08**	24.90	15.17

supports multiple humans but uses a temporally fixed, text-derived affordance graph and keeps human motions fixed during object optimization, leaving human errors uncorrected. MMHOI-Net[1] achieves favorable CD-h results, having been trained on CORE4D, although direct comparison is limited by insufficient reporting of evaluation details. It processes images independently without explicit temporal constraints on object trajectories or contacts, and its reported CD-o exceeds that of either variant of COMBO. In contrast, COMBO jointly refines human and object motions through a proximity-based temporal contact graph that couples the shared object to all interacting humans.

4.4 Ablation Studies

In Table 2, we ablate each component from the full pipeline of COMBO and examine the effect of the two optimization stages using the corresponding variants.

Effects of human temporal refinement. Since the initial human meshes and motions are estimated independently for each frame, omitting the temporal refinement in Eq. (1) introduces frame-wise jitter and degrades human-specific reconstruction metrics, *i.e.*, CD-h and PA-MPJPE. These temporal inconsistencies further affect the subsequent human-anchored depth calibration, propagating errors to object initialization and consequently degrading CD-o.

Effects of contact-based object optimization. Without contact-based object optimization (Stage 1), the human and object are directly optimized using Eq. (7). In this setting, the human body, particularly the articulated arm chain, can still benefit from the joint optimization by adjusting toward the object in the shared 3D space, resulting in relatively improved CD-h. In contrast, the object is optimized without a preceding coarse contact-guided refinement, which leads to degraded CD-o as well as larger object translation and rotation errors, measured by T_{obj} and R_{obj}.

Effects of joint human-object refinement. The joint human-object refinement (Stage 2) performs fine-grained refinement of the human-object configuration by encouraging contact consistency while preventing interpenetration. Although the independently evaluated poses of the human and object may not always improve, their relative alignment in the shared 4D space becomes more accurate, leading to better CD metrics for the full reconstruction. These results demonstrate that the two optimization stages play complementary roles: Stage 1 provides a reliable coarse object alignment based on contact, while Stage 2 further refines the joint human-object configuration. The proposed optimization stage curriculum therefore forms a complete optimization pipeline for collaborative human-object reconstruction.

4.5 COMBO Generalization to In-the-wild & Generated Videos

We demonstrate the generalization capability of our method by reconstructing collaborative HOIs from in-the-wild and synthetically generated videos. In Fig. 4, we generate videos (top two rows) using text prompts with diverse objects, *e.g.*, a duck-shaped inflatable tube and a wooden crate, and capture in-the-wild videos (bottom two rows) ourselves. We compare COMBO, CARI4D, and HOI-PAGE by rendering their reconstructed humans and

Table 3: **User study and VLM score on real/generated videos.** Higher is better.

Method	User score ↑		VLM score ↑	
	Real	Gen.	Real	Gen.
CARI4D	1.95	3.05	2.83	3.29
HOI-PAGE	2.05	2.19	3.50	3.05
COMBO (Ours)	**3.95**	**3.68**	**4.33**	**3.76**

[1]As the code of MMHOI-Net is not publicly available, we report the results on the CORE4D from its paper.

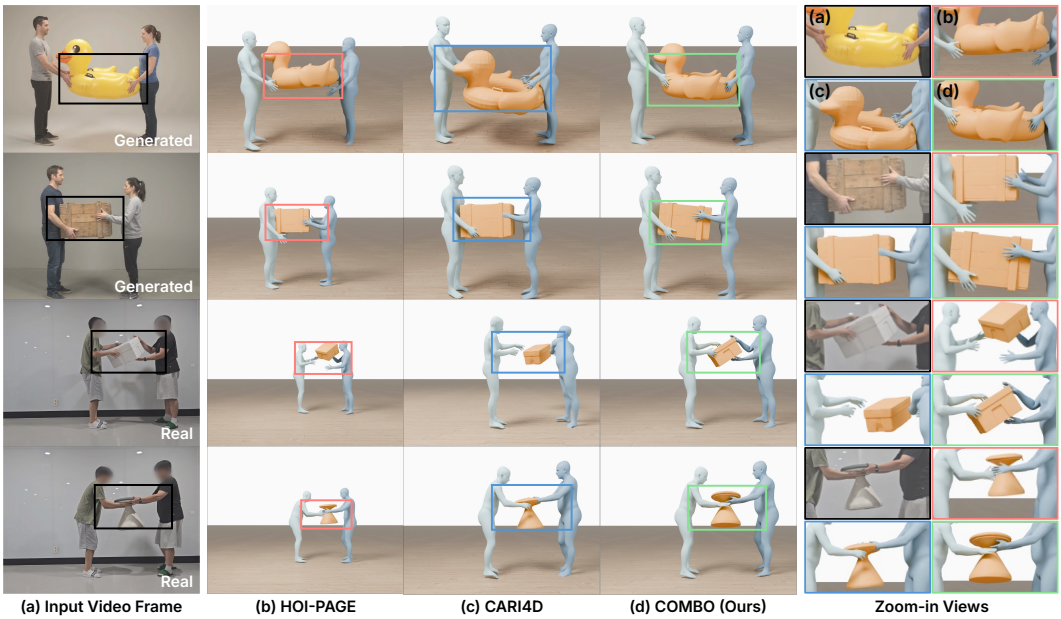

Figure 4: **Qualitative comparison on generated & in-the-wild videos.** We reconstruct collaborative 4D HOI from synthetically generated videos (top two rows) and in-the-wild videos (bottom two rows). COMBO demonstrates zero-shot generalization across diverse objects, with better image alignment and more plausible contact-aware reconstruction than competing methods.

objects with their predicted camera parameters. COMBO produces image-aligned reconstructions with plausible contacts and without severe mesh collisions, while CARI4D and HOI-PAGE exhibit configuration and scale misalignment, respectively. Since ground-truth annotations are unavailable for these videos, we quantitatively evaluate the motion quality with a user study and a VLM assessment, for details, refer to Sec. B.5. In Table 3, COMBO achieves the highest user and VLM scores on both in-the-wild and generated videos, demonstrating its strong generalization capability to diverse objects and unseen visual domains.

5 Conclusion and Limitations

Conclusion. We present COMBO, a training-free framework for reconstructing collaborative human-object interactions from monocular video. Without relying on predefined object templates or ground-truth camera parameters, our method reconstructs multiple humans and the interacting object in a shared 4D space. COMBO first establishes a human-anchored shared 4D space as a consistent reference for object reconstruction and optimization. COMBO subsequently couples all the human and object entities through a proximity-based temporal contact graph, which identifies which hand of which human contacts the object at each frame and captures contacts as they persist, break, and re-form over time. Experiments on a collaborative HOI benchmark, ablation studies, and evaluations on in-the-wild and generated videos show that COMBO yields more accurate and plausible reconstructions than existing methods. We hope COMBO serves as a scalable source of collaborative interaction data from everyday videos, complementing multi-view capture systems and supporting downstream applications such as human-robot collaboration.

Limitations. Although COMBO shows plausible motion reconstruction and generalization to in-the-wild and synthetically generated videos, it relies on 3D foundation models; therefore, a heavily occluded first frame or a lost track under prolonged occlusion may harm our reconstruction. However, our pipeline is also likely to benefit directly from future advances in these models. The reconstructed grasps may not capture detailed finger-object contacts, which could be improved by incorporating dedicated hand pose estimation models into our framework. Finally, since the temporal contact graph is built from human-object proximity, sequences in which the object moves freely without contact, *e.g.*, humans throwing a ball to each other, remain challenging.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks that require disclosure:

Support qualitative and thematic data analysis Generative AI tools assisted us in building the visualization framework used to inspect and compare results qualitatively, and in writing the scripts that compute our error metrics. The resulting comparisons and metrics were checked by the authors, and all interpretation of the outcomes is our own.

Implement the proposed method AI coding agents (Claude) assisted with parts of the training and evaluation scripts and with refactoring. All generated code was reviewed, verified, and tested by the authors; the core method design is our own.

We did not use generative AI tools to generate synthetic data, to clean or reformat datasets, to develop theoretical models or conceptual frameworks, to formulate or refine hypotheses, to design or give feedback on our research methodology or experiments, to assist with translation, or to interpret results. The other tasks that require disclosure (formulating mathematical claims, providing key steps of proofs, and writing proofs) do not apply here, since this work contains no formal theorems or proofs.

We did use generative AI tools for the following tasks, for which disclosure is recommended:

Create or modify scientific figures or images Generative AI tools helped automate the rendering pipeline used to produce result animations. However, we note that all source motion results in the paper were generated by our method.

Create or edit software code We used an AI coding agent (Claude) during development for auxiliary tasks, *e.g.*, writing training and evaluation scripts and refactoring code.

Edit a research paper to improve readability We used an LLM to review a draft of this paper and used its feedback to manually correct grammatical errors and improve clarity.

The authors reviewed all AI-assisted work. We read, verified, and tested all LLM-generated code for correctness, and every piece of AI-assisted text was either rewritten or explicitly approved by the authors. We take full responsibility for the final content of this paper, including any text, claims, or artifacts produced with the help of generative AI.

ETHICS STATEMENT

We disclose that all quantitative experiments are conducted on publicly available datasets and benchmarks, used in accordance with their licenses. For the in-the-wild videos, all participants were informed in advance that the captured videos would be used as experimental data in this paper, and we obtained their consent before recording. For the in-the-wild videos, to prevent exposure of the participants' identities, we blur their faces with mosaics in all videos and figures presented in the paper and the supplementary material.

REPRODUCIBILITY STATEMENT

The full pipeline of COMBO, including the human-anchored initialization, the proximity-based temporal contact graph, and the contact-guided collaborative refinement, is described in Sec. 3.1–3.4. The implementation details of each stage, including the pretrained models, hyperparameters, loss weights, and optimization schedules, are provided in Sec. A, and the runtime analysis is reported in Sec. A.4. The dataset construction, including the sequence selection criteria on CORE4D, and the evaluation protocol with the definitions of all reported metrics are documented in Sec. B.1 and Sec. B.3, respectively. The configurations used to run the competing methods, HOI-PAGE, CARI4D, and MMHOI-Net, are detailed in Sec. B.4. The text prompts for designing our user study and VLM-score evaluations are detailed in Sec. B.5. Finally, the capture and processing settings of the in-the-wild and generated videos are reported in Sec. B.2. We plan to publicly release the code upon the paper acceptance.

REFERENCES

Anthropic. Introducing Claude Sonnet 5, 2026. URL https://www.anthropic.com/news/claude-sonnet-5. 20

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026. 4, 14

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2, 5, 15

Yue Chen, Xingyu Chen, Yuxuan Xue, Anpei Chen, Yuliang Xiu, and Gerard Pons-Moll. Human3r: Everyone everywhere all at once. In *International Conference on Learning Representations (ICLR)*, 2025. 3

Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

Google DeepMind. Gemini Omni Flash: Model Card, 2026a. URL https://deepmind.google/models/model-cards/gemini-omni-flash/. 16

Google DeepMind. Veo 3.1 Lite: Model Card, 2026b. URL https://deepmind.google/models/model-cards/veo-3-1-lite/. 16

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 18

Dingbang Huang, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Recovering physically plausible human-object interactions from monocular videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3

Yinghao Huang, Omid Taheri, Michael J Black, and Dimitrios Tzionas. Intercap: Joint markerless 3d tracking of humans and objects in interaction from multi-view rgb-d images. *International Journal of Computer Vision (IJCV)*, 2024. 3

Rawal Khirodkar, He Wen, Julieta Martinez, Yuan Dong, Zhaoen Su, and Shunsuke Saito. Sapiens2. In *International Conference on Learning Representations (ICLR)*, 2026. 4, 14

Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. Mmhoi: Modeling complex 3d multi-human multi-object interactions. In *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2026. 2, 3, 19

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision (ECCV)*, 2024. 2

Lei Li and Angela Dai. HOI-PAGE: Zero-shot human-object interaction generation with part affordance guidance. In *International Conference on Machine Learning (ICML)*, 2026. 2, 3, 6, 18

Mengfei Li, Peng Li, Zheng Zhang, Jiahao Lu, Chengfeng Zhao, Wei Xue, Qifeng Liu, Sida Peng, Wenxiao Zhang, Wenhan Luo, et al. Unish: Unifying scene and human reconstruction in a feed-forward pass. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026. 2, 4

Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2, 3, 6, 8, 16

Zhizheng Liu, Joe Lin, Wayne Wu, and Bolei Zhou. Joint optimization for 4d human-scene reconstruction in the wild. In *International Conference on Learning Representations (ICLR)*, 2026. 2

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2015. 4

Lea Müller, Hongsuk Choi, Anthony Zhang, Brent Yi, Jitendra Malik, and Angjoo Kanazawa. Reconstructing people, places, and cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 2

Hyeongjin Nam, Daniel Sungho Jung, Gyeongsik Moon, and Kyoung Mu Lee. Joint reconstruction of 3d human and object via contact-based refinement transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

OpenAI. GPT-4o mini: Advancing Cost-Efficient Intelligence, 2024. URL https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. 18

YM Pradyumna, Yuxuan Xue, Yue Chen, Nikita Kister, István Sárándi, and Gerard Pons-Moll. Graft: Geometric refinement and fitting transformer for human scene reconstruction. In *European Conference on Computer Vision (ECCV)*, 2026. 2

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2017. 4

István Sárándi and Gerard Pons-Moll. Neural localizer fields for continuous 3d human pose and shape estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 4, 14

Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, and Michael J. Black. DECO: Dense estimation of 3D human-scene contact in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025a. 2

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026a. 2

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. 2

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *International Conference on Learning Representations (ICLR)*, 2026b. 2

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2022. 3

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. In *International Conference on 3D Vision (3DV)*, 2025. 3

Xianghui Xie, Bowen Wen, Yan Chang, Hesam Rabeti, Jiefeng Li, Ye Yuan, Gerard Pons-Moll, and Stan Birchfield. Cari4d: Category agnostic 4d reconstruction of human-object interaction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 2, 3, 7, 18, 19

Lixin Xue, Chengwei Zheng, Georgios Paschalidis, Chen Guo, Manuel Kaufmann, Juan Zarate, and Dimitrios Tzionas. RHINO: Reconstructing human interactions with novel objects from monocular videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 3

Pradyumna Yalandur-Muralidhar, Yuxuan Xue, Xianghui Xie, Margaret Kostyrko, and Gerard Pons-Moll. Physic: Physically plausible 3d human-scene interaction and contact from a single image. In *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2025. 2

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. Lemon: Learning 3d human-object interaction relation from 2d images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision (ECCV)*, 2020. 3

Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-m3: Capture multiple humans and objects interaction within contextual environment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

Ruihang Zhang, Felix Taubner, Pooja Ravi, Kiriakos N Kutulakos, and David B Lindell. Proxypose: 6-dof pose tracking via video-to-video translation. *arXiv preprint arXiv:2607.06555*, 2026. 5, 15

Yubo Zhao, Yujin Chai, Yunao Dong, Chengfeng Zhao, Zijiao Zeng, Yuan Liu, and Chi-Keung Tang. Real2sim in hoi: Toward physically plausible hoi reconstruction from monocular videos. *arXiv preprint arXiv:2605.14462*, 2026. 3

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. 15

COLLABORATIVE HUMAN-OBJECT INTERACTION RECONSTRUCTION FROM A MONOCULAR VIDEO — SUPPLEMENTARY MATERIAL —

Anonymous authors
Paper under double-blind review

This supplementary material provides implementation details and additional results that could not be included in the main paper due to space constraints. We also describe how the competing methods were configured for our experiments. We **encourage readers to watch the supplementary video** for dynamic visualizations of the reconstruction results.

CONTENTS

A Implementation Details **14**
A.1 Human-Anchored Initialization 14
A.2 Proximity-Based Contact Estimation 15
A.3 Contact-Guided Refinement 15
A.4 Runtime 16
B Experimental Setup **16**
B.1 Dataset 16
B.2 In-the-wild videos & generated videos 16
B.3 Evaluation Metrics 18
B.4 Competing Methods 18
B.5 User Study and VLM Evaluation 20
C Additional Results **21**
C.1 Additional Qualitative Results 21

A IMPLEMENTATION DETAILS

A.1 HUMAN-ANCHORED INITIALIZATION

Human motion refinement. We run NLF (Sárándi & Pons-Moll, 2024) using the person bounding boxes obtained from the SAM3 (Carion et al., 2026) masks and keep the predicted hand articulation fixed. The two terms of $\mathcal{L}_{\text{human}}$ are computed as:

$$\mathcal{L}_{\text{kpt}} = \sum_{i,t,j} w_{i,t,j} \left\| \Pi_{K_1}(\mathbf{J}_{i,t,j}) - \mathbf{j}_{i,t,j} \right\|^2, \qquad \mathcal{L}_{\text{acc}} = \sum_{i} \sum_{t=2}^{T-1} \left\| \mathbf{J}_{i,t-1} - 2\mathbf{J}_{i,t} + \mathbf{J}_{i,t+1} \right\|^2, \quad \text{(a)}$$

where $\mathbf{J}_{i,t,j}$ is the j-th joint regressed from $\mathbf{V}_{i,t}$, and $\mathbf{j}_{i,t,j}$ and $w_{i,t,j}$ are the corresponding Sapiens2 (Khirodkar et al., 2026) keypoint and its confidence among 63 body, foot, and hand keypoints. We optimize for 1,000 iterations with Adam (learning rate 10^{-3}), $\lambda_{\text{kpt}} = 10^{-2}$, and $\lambda_{\text{acc}} = 10^{4}$.

Depth calibration. We estimate the global depth scale and shift parameter a and b by a robust fit between the raw depth and the depth of the reference human mesh $\mathbf{D}_t^h$ over the pixels Ω_t inside its mask as:

$$\min_{a,b} \sum_t \sum_{\mathbf{u}\in\Omega_t} \left| a\,\mathbf{D}_t^{\text{raw}}(\mathbf{u}) + b - \mathbf{D}_t^h(\mathbf{u}) \right|. \tag{b}$$

Object initialization. SAM3D (Chen et al., 2026) takes the first frame, its object mask, and the point map back-projected from $\mathbf{D}_1$. We align its output in four steps: we (i) move the object along the viewing ray to match $\mathbf{D}_1$, (ii) refine its rotation by registering the visible surface to $\mathbf{X}_1$ with Iterative Closest Point (ICP), (iii) align its upright axis with the floor normal estimated from the static foot joints of the humans, and (iv) select the scale in $[0.85, 1.2]$ that maximizes the silhouette IoU. From the scaled mesh $\mathbf{O}$, we sample 20,000 surface points $\mathcal{P}$ and precompute its signed distance field SDF.

Object motion initialization. Since ProxyPose (Zhang et al., 2026) processes fixed-length clips and predicts the motion relative to the first frame of each clip, we run it on 49-frame clips overlapping by one frame and chain the relative rotations,

$$\mathbf{Q}_{1\to t} = \mathbf{Q}_{t_0\to t}\,\mathbf{Q}_{1\to t_0}, \qquad \mathbf{R}_t^{\text{init}} = \mathbf{Q}_{1\to t}\mathbf{R}_1, \tag{c}$$

where t_0 is the first frame of the clip containing t. The translation places the centre of $\mathcal{P}$ at the center of the depth points by:

$$\mathbf{t}_t^{\text{init}} = c(\mathbf{X}_t^{\text{fit}}) - \mathbf{R}_t^{\text{init}}\,c(\mathcal{P}), \tag{d}$$

where $c(\cdot)$ is the bounding-box center and $\mathbf{X}_t^{\text{fit}} \subset \mathbf{X}_t$ excludes the points on the mask boundary and in the background.

A.2 Proximity-Based Contact Estimation

The distances in Eq. (3) are

$$d_{i,s,t}^{\text{2D}} = \min_{\mathbf{v}\in\mathbf{V}_{i,s,t}} \text{DT}_{\mathbf{m}_t}\big(\Pi_{K_t}(\mathbf{v})\big), \qquad d_{i,s,t}^{\text{3D}} = \min_{\mathbf{v}\in\mathbf{V}_{i,s,t}} d(\mathbf{v}, \mathbf{X}_t), \tag{e}$$

where $\text{DT}_{\mathbf{m}_t}$ is the distance transform of the object mask in pixels. We use $\sigma_{\text{2D}} = 30$ pixels, $\sigma_{\text{3D}} = 0.08\,\text{m}$, $\gamma = 0.1$, and a window $\mathcal{W}_t$ of 9 frames.

A.3 Contact-Guided Refinement

We parameterize $\mathbf{R}_t$ with the continuous 6D representation (Zhou et al., 2019), write the posed surface points as $\mathcal{P}_t = \mathbf{R}_t\mathcal{P} + \mathbf{t}_t$, and denote a human vertex in the object frame by $\bar{\mathbf{v}}_t = \mathbf{R}_t^\top(\mathbf{v}_t - \mathbf{t}_t)$.

Contact term. Since a hand touches the object with only part of its surface, we average over the hand vertices $\mathcal{N}_{i,s,t}$ closest to the object (10% of $\mathbf{V}_{i,s,t}$),

$$d_{\text{cont}}(\mathbf{V}_{i,s,t}, \mathbf{O}_t) = \frac{1}{|\mathcal{N}_{i,s,t}|} \sum_{\mathbf{v}\in\mathcal{N}_{i,s,t}} d(\mathbf{v}, \mathcal{P}_t)^2. \tag{f}$$

Observation and prior terms. $\mathcal{L}_{\text{init}}$ consists of a depth term, a mask term, and a pose prior,

$$\mathcal{L}_{\text{3D}} = \sum_t \rho\Big(\sum_{\mathbf{x}\in\mathbf{X}_t^{\text{fit}}} d(\mathbf{x}, \mathcal{P}_t)^2 \Big), \tag{g}$$

$$\mathcal{L}_{\text{2D}} = \sum_t \sum_{\mathbf{u}\in\mathbf{m}_t} d\big(\mathbf{u}, \Pi_{K_t}(\mathcal{P}_t)\big)^2, \tag{h}$$

$$\mathcal{L}_{\text{prior}} = \sum_t \|\mathbf{t}_t - \mathbf{t}_t^{\text{init}}\|^2 + \angle(\mathbf{R}_t, \mathbf{R}_t^{\text{init}})^2. \tag{i}$$

where ρ is the Geman-McClure function, $\mathbf{u}$ is a mask pixel, and $\angle(\cdot,\cdot)$ is the geodesic distance.

Temporal term. $\mathcal{L}_{\text{temp}}$ smooths the translation correction $\boldsymbol{\delta}_t = \mathbf{t}_t - \mathbf{t}_t^{\text{init}}$ and keeps each contacting hand static in the object frame,

$$\mathcal{L}_{\text{smooth}} = \sum_t \left\| \boldsymbol{\delta}_t - \tfrac{1}{2}(\boldsymbol{\delta}_{t-1} + \boldsymbol{\delta}_{t+1}) \right\|^2, \qquad \mathcal{L}_{\text{drift}} = \sum_t \sum_{\mathbf{v}} \|\bar{\mathbf{v}}_{t+1} - \bar{\mathbf{v}}_t\|^2, \tag{j}$$

where $\mathcal{L}_{\text{drift}}$ sums over the hand vertices of the active edges.

Penetration term. $\mathcal{L}_{\text{pene}} = \sum_t \sum_{\mathbf{v}} \max\big(0, -\text{SDF}(\bar{\mathbf{v}}_t)\big)$ penalizes the body vertices inside the object.

Optimization. Both refinement steps run for 600 iterations with Adam (learning rate 10^{-2}), and the joint refinement starts from the object-only result. The loss weights are 10^3 for $\mathcal{L}_{\text{cont}}$, $\mathcal{L}_{\text{3D}}$, and $\mathcal{L}_{\text{smooth}}$, 10^2 for $\mathcal{L}_{\text{prior}}$, and 10^{-4} for $\mathcal{L}_{\text{2D}}$. The weights of $\mathcal{L}_{\text{drift}}$ and $\mathcal{L}_{\text{pene}}$ increase linearly over the iterations, from 10 to 10^3 and from 0 to 10, respectively. In the joint refinement, $\mathcal{L}_{\text{human}}$ uses the same weights as in the human motion refinement.

A.4 Runtime

We measure the runtime on a single NVIDIA RTX A6000 GPU. Processing a CORE4D sequence of median length (70 frames) takes approximately 40 minutes. Most of the time is spent on the object reconstruction and alignment (8 minutes), the object tracking (8 minutes), and the two refinement steps (16 minutes), whereas the contact estimation takes only a few seconds.

B Experimental Setup

B.1 Dataset

We evaluate on CORE4D-Real (Liu et al., 2025), where two humans manipulate a shared object. From the sequences of the S1 test split, we extract the segments in which the humans carry the object, select for each sequence the camera view where the object and both humans are most visible, and exclude the sequences with thin objects, insufficient visibility, or inconsistent ground truth. This results in 150 sequences with 10,930 frames at 15 fps, covering five object categories: desk, bucket, box, chair, and board. For the ablation study, we use 100 of these sequences selected by stratified sampling over the object category and the sequence length.

B.2 In-the-wild videos & generated videos

In-the-wild videos. We capture videos of two humans manipulating everyday objects, such as chairs, desks, trash cans, and boxes, with a Samsung Galaxy S23 at 1920×1080 resolution and about 30 fps. The videos are converted to a constant frame rate of 30 fps before processing. We use the same pipeline and hyperparameters as for CORE4D, with the object category as the SAM3 text prompt, *e.g.*, "chair", "table", "stool", "trash can", and "wheeled cart".

Generated videos. We generate videos at 1280×720 resolution and 24 fps using Veo3.1 Lite (Google DeepMind, 2026b), and Gemini Omni 1.1 Flash (Google DeepMind, 2026a). For each video, we provide the object category as the SAM3 text prompt, *e.g.*, "cardboard box", "wooden crate", "inflatable duck", "log", "piano", and "sofa". When the object mask is split into multiple tracks, we merge the tracks that do not overlap in time, and we trim the frames in which the object leaves the image. The video generation prompt varied across samples. Below, we provide the prompt used for the wooden-crate handover example.

Video Generation Prompt Example (Wooden-crate Handover)

```
Generate a realistic video for use in our experimental pipeline.

[Scenario]

Exactly two people and one large wooden crate are fully visible in the initial frame. The
first person carries the crate sideways toward the second person. The second person then
grasps the opposite side of the crate, and both people hold it together briefly during the
```

```
 handover. After that, the first person releases the crate, and the second person carries
it for several more steps before placing it down at a new location.

Requirements

1. Strictly Fixed Camera

- The camera must remain completely stationary and locked in place for the entire video.
- Use one continuous wide shot from a tripod-like fixed viewpoint.
- No camera movement, tracking, panning, tilting, zooming, dolly movement, rotation,
reframing, stabilization-induced movement, or viewpoint change.
- The camera must NEVER move or reframe to keep the people or crate visible.
- The first frame and the last frame must use exactly the same camera position,
orientation, zoom, and field of view.

2. Movement Must Stay Inside the Fixed Frame

- Both people and the wooden crate must already be fully visible in the initial frame.
- Start the first person and the crate on one side of the frame and position the second
person farther along the direction of travel.
- The first person must take several clearly visible walking steps while carrying the
crate toward the second person.
- The handover must occur clearly within the original fixed camera frame.
- The crate must remain lifted off the ground during the carrying and handover phases.
- The starting position and final position of the wooden crate must be clearly different.
- All movement must remain within the boundaries of the original fixed camera frame.
- The camera must not move, pan, zoom, or reframe to follow the subjects.

3. Large and Rigid Object

- Use one large wooden crate with simple, clearly recognizable rectangular geometry.
- Use a solid box-shaped wooden crate with sturdy construction and minimal surface details.

- The crate should be large and bulky enough that carrying and handing it over is clearly
visible.
- The wooden crate must remain a single rigid object with a stable and consistent shape
throughout the entire video.
- Its overall geometry, dimensions, proportions, size, silhouette, edges, corners, and
structural details must remain unchanged from the beginning to the end of the clip.
- Do not deform, bend, stretch, squash, warp, wobble, twist, morph, or resize the crate.
- Do not add, remove, duplicate, or alter any structural components while the crate is
being carried or transferred.
- The wooden crate must preserve exactly the same overall shape and appearance in every
frame.

4. Clear Handover

- The transfer from the first person to the second person must be clearly visible.
- The second person must establish a firm grasp on the crate before the first person
releases it.
- Both people must simultaneously hold the crate for a short moment during the transfer.
- The first person must visibly release both hands after the second person has taken
control.
- The crate must not touch the ground during the handover.

5. Minimal Occlusion

- Keep both people and the wooden crate clearly visible from the camera viewpoint.
- Arrange the interaction so that both people's hands and the crate are clearly visible
during the transfer.
- Avoid one person fully blocking the other.
- Avoid severe human-object occlusion.

6. Subject Visibility

- Both people should remain fully visible from head to feet throughout the interaction.
- The entire wooden crate should remain visible throughout the interaction.
- Visibility must be achieved through subject placement and movement within the fixed
frame, not through camera movement.

7. Exactly Two People Only

- The video must contain EXACTLY TWO people in total.
- Only these two people may appear anywhere in the video.
- Do not generate additional people, background people, partially visible people,
reflections of other people, or people entering or leaving the scene.
- Do not duplicate either person.
- Do not create extra human figures, limbs, bodies, or silhouettes.
```

```
8. Simple Background

- Use a very simple, minimal indoor environment with a plain wall and an empty floor.
- Keep the background mostly empty.
- The background must contain no other people.
- Use consistent, neutral lighting.

9. Visual Consistency

- The wooden crate must remain visually consistent across all frames.
- Do not allow any temporal inconsistency in the object's shape or structure.
- The crate must behave like a heavy, solid, rigid wooden object with stable geometry.
- No frame-to-frame shape changes, size changes, part duplication, part disappearance, or
object morphing are allowed.
- Both people must remain visually consistent throughout the video without duplication,
disappearance, or identity changes.
```

B.3 Evaluation Metrics

Following CARI4D (Xie et al., 2026), we align each prediction to the ground truth with a single similarity transform estimated at the first frame and apply it to the entire sequence. To extend this protocol to multiple humans, we match the predicted and ground-truth humans in the first frame using their projected 2D joints, and estimate a single transform jointly from the vertices of all humans, thereby preserving errors in their relative placement. CD-h, CD-o, and CD-c are the bidirectional Chamfer distances in centimeters between 8,196 points sampled on the surfaces of each human mesh, object mesh, and the combined meshes of all humans and the object, respectively. PA-MPJPE is the error of the 22 SMPL body joints in centimeters after per-frame Procrustes alignment of each matched human. T_{obj} is the distance between the object centroids in centimeters, and R_{obj} is the geodesic distance between the object rotations relative to the first frame in degrees. The human metrics are averaged over the matched humans, and all metrics are averaged over all frames.

B.4 Competing Methods

B.4.1 HOI-PAGE

HOI-PAGE (Li & Dai, 2026) is a zero-shot method that generates 4D human-object interactions from a text prompt. It queries a large language model with the prompt and the object information to construct a part affordance graph, which specifies the contacts between the human and object parts over the entire sequence, and optimizes the object motion to follow a generated reference video under these contact constraints. To adapt it to reconstruction, we omit its video diffusion stage and use the RGB videos of CORE4D as the reference videos. As the text prompt, we generate a caption for each video with GPT-4o mini (OpenAI, 2024), which receives the chronologically ordered frames, the CORE4D action label, and the object name, and refers to the two participants as "Person 1" and "Person 2" consistently with the action labels. Following the demo code of HOI-PAGE, we use DeepSeek-R1 (Guo et al., 2025) to propose four candidate initial rotations of the object about its y-axis, allowing up to three API calls per query and falling back to the candidates $\{0°, 90°, 180°, 270°\}$ when all calls fail. The action label definitions, the captioning prompt, and example captions are listed below.

Caption generation details.

CORE4D Action Label Definitions

```
ACTION_LABEL_DEFINITIONS = {
"move1": "The two people move the object together; only one person knows the target
position.",
"move2": "The two people move the object together; both people know the target position.",
"raise": "The two people raise the object together.",
"raise1": "The two people raise the object together.",
"raise2": "The two people raise the object together.",
"raise3": "The two people raise the object together.",
"rot": "The two people rotate the object together by 180 degrees.",
"raise_rot": "The two people raise the object and rotate it together by 180 degrees.",
"raise1_rot": "The two people raise the object and rotate it together by 180 degrees.",
```

```
"raise2_rot": "The two people raise the object and rotate it together by 180 degrees.",
"raise3_rot": "The two people raise the object and rotate it together by 180 degrees.",
"pass1": "Person 1 passes the object to Person 2.",
"pass2": "Person 2 passes the object to Person 1.",
"join": "Person 2 joins the object interaction halfway through.",
"leave": "Person 2 leaves the object interaction halfway through.",
"strike": "One person waves the object in their hands while the other person dodges it.",
"notify": "One person sits on the object, then the other person notifies them and moves
the object away.",
}
```

Prompt for GPT-4o mini

```
DEFAULT_PROMPT = """
You caption short human-object interaction videos for an action dataset.
The images are chronologically ordered frames from one video.
The user message provides authoritative action and object metadata. Follow that
metadata when it resolves visual ambiguity, but do not output the label code.
Write exactly one detailed but concise sentence describing the main action.
The caption must name exactly one physical object: the authoritative object name
provided by the user. Include that exact singular noun at least once. Never name,
describe, compare with, or refer to any other physical object, including
background furniture, containers, destinations, or obstacles. Do not replace the
authoritative object noun with a visually inferred alternative.
Use the action metadata to choose the action wording. For every two-person
interaction, explicitly name both dataset identities as "Person 1" and
"Person 2". Never use anonymous alternatives such as "both people", "two
people", "the people", or "they". When they do the same action, write
"Person 1 and Person 2 ..."; when their roles differ, write "Person 1 ... while
Person 2 ...". These identities are defined by the action label, not by screen
position.
When a person's hand contact is clearly visible, describe it specifically as
"both hands", "left hand", "right hand", or "one hand" for that person. Omit
hand-use wording when the frames do not make it clear. Do not claim hidden
knowledge, intent, or other details not visible in the video.
Return the caption only, with no labels, quotation marks, or explanation."""
```

Example Captions

```
"Person 1 and Person 2 move the box together towards the target position."
"Person 2 passes the desk to Person 1."
"Person 1 and Person 2 rotate the desk together by 180 degrees."
"Person 1 holds the board while Person 2 joins to assist in handling the board."
```

B.4.2 CARI4D

CARI4D (Xie et al., 2026) reconstructs category-agnostic 4D interactions between a single human and an object from a monocular video. It reconstructs the object mesh from the input video, predicts frame-wise contacts for the two hands of the human, and refines the interaction with these contacts at metric scale. Since CARI4D handles a single human per run, we run it twice on each CORE4D video, once per human, with the corresponding SAM3 human mask, while both runs share the same SAM3 object mask and canonical object mesh. We follow the inference and optimization settings of its official custom video demo and use the publicly released pretrained model without additional training. Both runs produce reconstructions in the input camera coordinates, so we combine their outputs directly without additional alignment. We keep both human reconstructions and fuse the two object estimates: the object size is set to the mean 3D bounding-box diagonal of the two metric object meshes, and at each frame, the translations are averaged and the rotations are combined by the chordal L_2 mean on $\mathrm{SO}(3)$. No further optimization is performed after the fusion, and the fused object poses and scale are used for both evaluation and visualization.

B.4.3 MMHOI-NET

MMHOI-Net (Kogashi et al., 2026) reconstructs multiple humans interacting with objects from a single image, using pre-scanned object templates, and is trained on CORE4D. Since its code is not

publicly available, we report the numbers from its paper on the CORE4D test set, obtained with its multi-human alignment. The paper reports only CD-h and CD-o on CORE4D, so the other metrics are not available. Note that these numbers are obtained on the test split and with the evaluation protocol of MMHOI-Net, which differ from ours, and are provided as a reference.

B.5 User Study and VLM Evaluation

User study. We conduct a user study with 30 participants on nine samples: two in-the-wild videos and seven generated videos. For each sample, participants are shown the reference RGB video and three anonymized reconstruction results from COMBO, CARI4D, and HOI-PAGE. The methods are randomly assigned to labels A, B, and C for each sample. Participants independently rate how accurately each result reproduces the reference on a 1–5 scale (1: very inaccurate; 5: very accurate). They are instructed to consider human motion, object motion, hand–object contact, and the progression of the interaction, prioritizing motion and interaction over appearance. We average the scores over participants and samples and report results for in-the-wild and generated videos separately.

VLM evaluation. We conduct a blinded VLM-based evaluation using Claude Sonnet 5 (Anthropic, 2026) on the nine samples as in the user study. For each sample, we uniformly select eight time points over the shared video duration and render each as a 2×2 comparison image containing the reference RGB frame and three anonymized reconstruction results. We randomly assign the three methods to labels A, B, and C independently for each sample, while keeping the assignment fixed across its eight frames. Given the eight chronologically ordered comparison images, the VLM independently rates A, B, and C according to their fidelity to the reference in terms of human motion, object motion, hand–object contact, and interaction progression, using a 1–5 scale (1: very inaccurate; 5: very accurate). We repeat each sample-level evaluation three times, average the three scores for each method and sample, and then average across samples. We report results for generated and in-the-wild videos separately. The prompt used for VLM evaluation is provided below.

Prompt for VLM Evaluation

```
You are an independent evaluator of human-object interaction video fidelity.

For one sample, you will receive eight 2x2 comparison images in chronological order. In
every image, the upper-left panel, `Reference`, is the original video at that time point.
The upper-right `A`, lower-left `B`, and lower-right `C` panels are three anonymized
results. You do not know which methods produced A, B, or C; do not try to infer them.

Independently assess how accurately A, B, and C reproduce the Reference. Focus on:
- the person's motion and pose changes;
- the object's motion and position changes;
- hand-object contact and manipulation relationships; and
- how the interaction progresses over time.

Prioritize motion and interaction accuracy over appearance or texture similarity. Judge
only what is visible in the images; do not infer unobserved or off-screen actions. Do not
rank A, B, and C against one another. Instead, assign each an independent score against
the absolute criteria below.

Score scale:
- 1: Very inaccurate
- 2: Inaccurate
- 3: Moderate
- 4: Accurate
- 5: Very accurate

Return only the JSON object below. Each score must be an integer from 1 to 5. For each `
rationale`, write one or two concise English sentences based on the observed temporal
motion and interaction evidence.

{
  "A": {"score": 1, "rationale": "..."},
  "B": {"score": 1, "rationale": "..."},
  "C": {"score": 1, "rationale": "..."}
}
```

C ADDITIONAL RESULTS

Comparison setup. We run HOI-PAGE and CARI4D as described in Secs. B.4.1 and B.4.2, respectively. For visualization, we render the reconstructed humans and object using the camera parameters predicted by the corresponding method. We use the same rendering settings for all methods and do not manually adjust the predicted cameras.

C.1 ADDITIONAL QUALITATIVE RESULTS

In Fig. S1, we provide additional qualitative comparisons with HOI-PAGE and CARI4D on CORE4D.

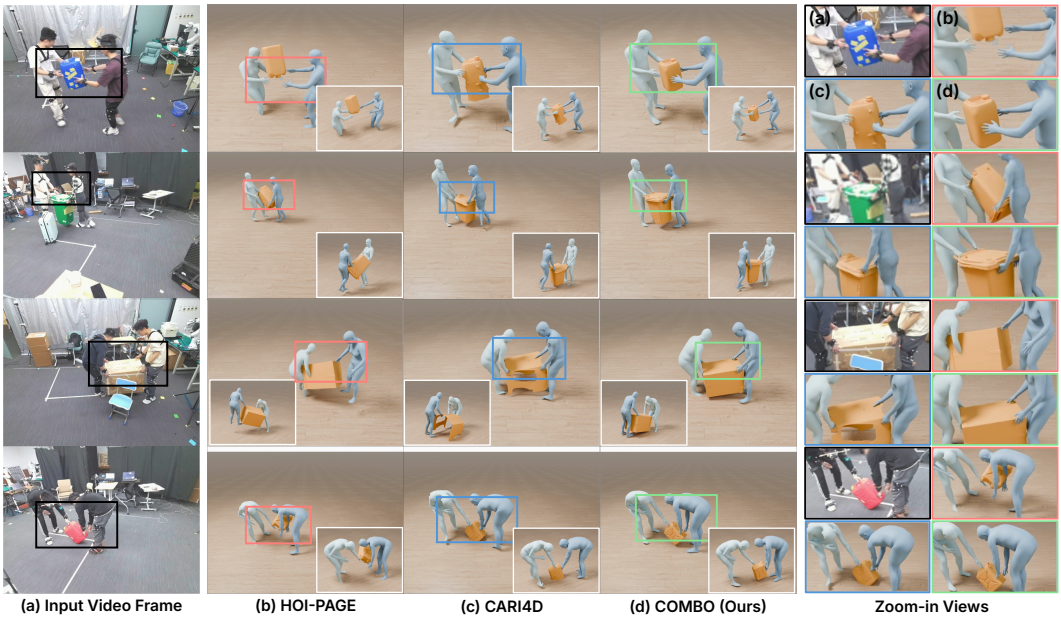

Figure S1: **Additional Qualitative comparison on CORE4D.** Visualization of our reconstruction results compared with competing methods in camera-view and side-view renderings (insets). The motions reconstructed by COMBO are well aligned with input video frames while maintaining plausible contacts and spatial configurations between multiple humans and the shared object.