4DHOISOLVER: EFFICIENT AND ACCURATE HUMAN–OBJECT MOTION RECONSTRUCTION FROM MONOCULAR VIDEOS

Anonymous authors
Paper under double-blind review

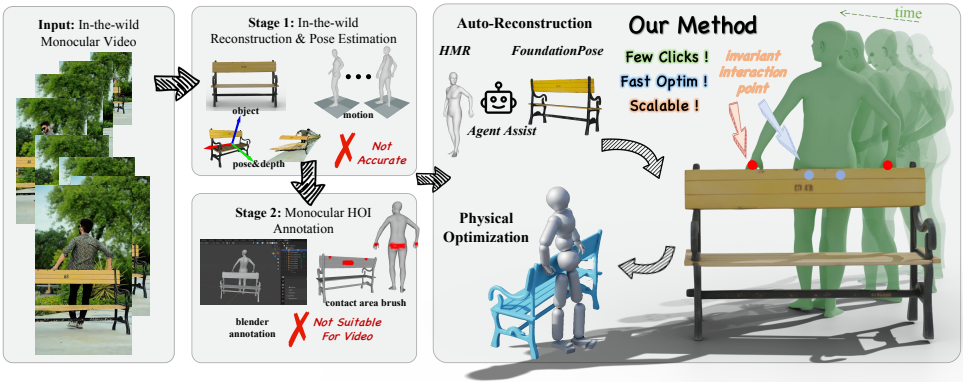


Figure 1: Existing monocular reconstruction is vulnerable to pose and depth errors, while frame-wise manual correction is difficult to scale to video. 4DHOISolver combines agent-assisted automatic reconstruction with sparse, temporally propagated point constraints for efficient repair, followed by unified physics-based refinement to recover coherent and physically plausible HOI motion.

ABSTRACT

Internet videos capture a rich diversity of human activities, objects, and environments, offering a readily available source of interaction data for embodied intelligence. However, recovering accurate and physically plausible 4D human–object interactions (HOI) from these monocular videos remains challenging due to depth ambiguity, occlusion, and complex contact relationships. We introduce **4DHOISolver**, a pipeline that combines automatic reconstruction with efficient manual correction. Our agent-based reconstruction method uses semantic interaction reasoning to guide geometric optimization, but challenging in-the-wild videos can still lead to reconstruction failures. These failures motivate an efficient and scalable manual repair method that corrects reconstructed motion through simple, sparse point annotations. We further propose a unified physical refinement method that corrects residual contact errors and interpenetration across diverse interactions without action-specific reward design.

1 INTRODUCTION

The quest for generalized robot systems, capable of understanding and interacting with the complex real world, fundamentally relies on rich, high-fidelity 4D (3D + time) human-object interaction (HOI) data. This data is invaluable for providing crucial insights into human motor skills, intent, and physical reasoning, empowering applications ranging from training intelligent robotic agents for dexterous manipulation to populating virtual worlds with realistic digital humans for gaming and VR/AR applications.

Despite its importance, acquiring generalized 4D HOI data remains a significant challenge. Existing multi-view HOI capture systems, exemplified by BEHAVE Bhatnagar et al. (2022), recover human and object geometry using calibrated RGB-D cameras. However, the associated cost and capture requirements limit large-scale data collection. Recognizing the limitations of high-precision cap-

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.

ture systems, recent research has explored reconstructing 3D HOI from widely available monocular images/videos. CARI4D Xie et al. (2026) enables automatic, category-agnostic 4D HOI reconstruction from monocular internet videos, demonstrating the potential of automatic methods beyond controlled capture. Nevertheless, in-the-wild HOI reconstruction remains highly complex, and automatic methods often fail on challenging cases. Pose and depth alignment errors can yield physically inaccurate contacts, motivating explicit human guidance. For example, methods like Open3DHOI Wen et al. (2025) rely on manual object pose adjustment in Blender, while PICO Cseke et al. (2025) proposes annotating contact regions. For videos, both require laborious annotations that are difficult to propagate across frames.

To address these critical challenges, we introduce **4DHOISolver** (Fig. 1), a framework that combines automatic reconstruction with efficient and high-quality manual correction. Our agent-based automatic reconstruction translates semantic reasoning about interaction phases, contact regions, and scene attributes into constraints for geometric optimization. To repair errors that remain in automatic results, we propose an efficient and scalable repair method driven by sparse point correspondences between the human, object, and video observations. These annotations provide explicit contact and projection constraints for a two-stage solver: rapid geometric alignment through least-squares matching, followed by gradient-based joint optimization to refine interaction geometry and temporal consistency.

Geometric correction alone does not ensure physical plausibility, as repaired trajectories may still contain interpenetration or inaccurate contacts. We therefore propose a physics-based refinement method that uses reference motion and contact annotations to guide motion imitation in simulation. The same reward formulation and training procedure apply across interaction types, without requiring action-specific reward design.

Overall, our work includes the following contributions:

- An agent-based automatic HOI reconstruction method that translates semantic interaction reasoning into structured constraints for geometric optimization.
- An efficient and scalable manual repair method that uses sparse contact and projection annotations to optimize HOI reconstructions through least-squares pose recovery, temporal fusion, and joint refinement.
- A unified interaction-agnostic physical refinement method for diverse interactions that improves physical plausibility.

2 Related Works

2.1 Monocular HOI Reconstruction

Dedicated HOI capture systems use multi-view RGB-D cameras Bhatnagar et al. (2022); Huang et al. (2024); Liu et al. (2025), large camera arrays Zhang et al. (2023; 2024), or Mocap suits Li et al. (2023); Lu et al. (2025); Jiang et al. (2024). Their hardware and controlled settings limit the diversity of captured interactions. Recent work reconstructs HOI from widely available monocular images and videos Yang et al. (2024); Dwivedi et al. (2025); Jiang et al. (2025); Li et al. (2025a); Geng et al. (2026); Xue et al. (2026); Wang et al. (2026); Li et al. (2026). Within this setting, methods handle objects by optimizing template keypoints Huo et al. (2024), manually refining generated meshes Wen et al. (2025), or retrieving proxy meshes and imposing contact constraints Cseke et al. (2025). CARI4D Xie et al. (2026) further enables automatic, category-agnostic 4D HOI reconstruction from monocular video through foundation-model predictions and joint refinement.

2.2 3D Reconstruction Tools

Robust 3D reconstruction of humans and objects is fundamental to recovering HOI from monocular videos. For human motion, while methods Lin et al. (2023); Cai et al. (2023); Sun et al. (2024) offer unified body-hand-face recovery, they often lack global trajectory awareness. Global human motion estimators Shen et al. (2024); Wang et al. (2024); Shin et al. (2024); Ye et al. (2023); Goel et al. (2023); Li et al. (2025b) instead recover body motion aligned with the camera or world frame, while

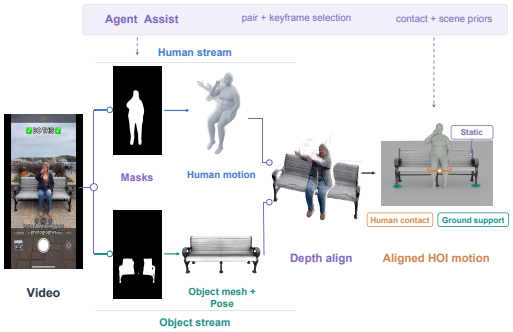


Figure 2: Agent-assisted automatic reconstruction, where semantic reasoning guides input selection and interaction-aware optimization to recover a complete 4D HOI trajectory.

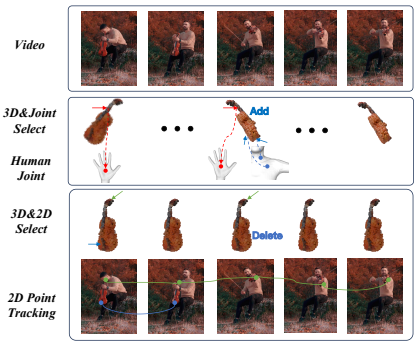


Figure 3: Sparse point annotation interface. Users specify 3D human–object contacts and object-to-image correspondences, supported by temporal 2D point tracking.

dedicated hand methods Pavlakos et al. (2024); Potamias et al. (2025); Zhang et al. (2019); Chen et al. (2025); Yu et al. (2025b) provide the fine-grained hand poses needed for interaction analysis.

For object reconstruction, recent 3D generative models Xiang et al. (2025b;a); Shi et al. (2023); Xu et al. (2024); Yu et al. (2025a) trained on massive datasets Deitke et al. (2023) can now produce high-quality meshes from real-world images. Furthermore, methods targeting real "in-the-wild" images, such as SAM 3D Team et al. (2025), enable object 3D reconstruction even in the presence of occlusions.

2.3 Physics-Based HOI Refinement

Physics-based motion imitation commonly formulates character control as a reinforcement learning problem, where a policy tracks kinematic reference motion while obeying simulated dynamics. DeepMimic Peng et al. (2018) established example-guided imitation for dynamic character skills, while SimPoE Yuan et al. (2021) and PHC Luo et al. (2023) extended physics-based control to human motion recovery and large-scale full-body motion tracking.

Physics-based HOI methods further incorporate objects and contact into motion imitation. PhysHOI Wang et al. (2023) introduces contact-graph supervision for dynamic interactions, SkillMimic Wang et al. (2025) learns object-centric basketball skills, and InterMimic Xu et al. (2025) scales whole-body control across diverse objects and interactions. InterPrior Xu et al. (2026) further explores generative control for HOI, while RePHO Huang et al. (2026) targets the physical refinement of noisy kinematic reconstructions from monocular videos.

3 Method

4DHOISolver first reconstructs human and object motion from monocular video with agent-assisted initialization (Sec. 3.1). Users then annotate sparse contact and object-to-image correspondences (Sec. 3.2), which guide temporal pose repair (Sec. 3.3). The complete workflow is illustrated in Appendix Fig. 9.

3.1 Agent-Based Automatic Reconstruction

Our pipeline reconstructs human and object motion from video and jointly refines their interaction (Fig. 2). A vision–language agent provides semantic interaction guidance throughout the reconstruction process.

3.1.1 Automatic Reconstruction

1) Instance selection and geometry. Starting from the input video, GroundingDINO Liu et al. (2023b) detects candidate people and objects, from which the agent identifies the main interacting

pair. SAM2 Ravi et al. (2024) then propagates their masks bidirectionally throughout the sequence. Based on object visibility and completeness, the agent further selects a suitable reconstruction frame for SAM 3D Objects Team et al. (2025) to recover the object mesh. In parallel, GENMO Li et al. (2025b) reconstructs global body motion, which HaMeR Pavlakos et al. (2024) complements with detailed hand poses, yielding a full-body human initialization in the world coordinate system.

2) Object tracking and spatial alignment. We estimate scene depth from the monocular RGB frames and calibrate it against the rendered human depth to establish a shared camera-space scale. FoundationPose Wen et al. (2024) provides the initial object pose, after which we robustly align the human root and object translation to the calibrated depth surfaces while preserving the object rotation and scale. Confidence-weighted temporal smoothing further stabilizes unreliable or occluded frames. The resulting trajectories are transformed into a shared world coordinate system to initialize joint HOI optimization. Implementation details are provided in Sec. A.1.

3.1.2 Agent-Assisted HOI Optimization

Agent input and prior prediction. We use Qwen3-VL-8B-Instruct Bai et al. (2025) to analyze temporally sampled full-scene frames and object-centered crops. It predicts the interaction type, contact phases, interacting body regions, and object-motion and ground-support priors. Input formatting is detailed in Sec. A.1.

Contact-guided optimization During predicted contact phases, we reduce the surface distance between the object and the relevant human region. Only the associated kinematic chain is refined, with temporal regularization to maintain smooth motion. This corrects missing contacts without manually specifying contact frames or body parts.

Scene-guided optimization Static-object priors encourage a consistent world-space pose. For ground-supported objects, we align estimated support regions with the ground plane while preventing penetration; dynamic objects are not subject to these static-scene constraints. Thus, the agent specifies semantic relationships that the optimizer enforces geometrically.

3.2 Sparse Point Annotation

Automatic reconstruction can leave interaction errors that require explicit geometric guidance. Our point-guided repair method corrects contact misalignment, depth ambiguity, and object-pose errors through sparse annotations. These annotations provide direct geometric control over challenging interactions, enabling efficient motion repair with limited manual effort while maintaining temporal consistency.

3.2.1 Contact Definition

A key prerequisite for annotating HOI point pairs is to define the interaction keypoints. Previous work predefined keypoints on object templates, which is not suitable for open-world object reconstruction. We choose to divide human joints as finely as possible while freely selecting the corresponding points on the object. We adopt a *tree structure* to define human keypoints, where **26** main body parts serve as parent nodes, and the child nodes represent finer subdivisions of each part, a total of **87** keypoints. For example, we divide the forearm into four points—front, back, left, and right—to ensure that contacts from all directions have corresponding interaction points.

3.2.2 Annotation Procedure

To annotate HOI motion efficiently, we built an annotation app, as shown in Fig. 3. Users provide two types of correspondences. For human–object contact, they select a 3D object point p_i^{3D} and the corresponding human joint q_i^{3D} in the joint tree. For object-to-image alignment, they select a 3D object point p_i^{2D} and click its corresponding 2D image location q_i^{2D}.

Many HOI sequences contain contact points that remain stable across multiple frames. We track these points to reduce annotation effort, and users re-annotate only when the contact changes. Specifically, we use a point-tracking model Karaev et al. (2025) to track the annotated 2D points. The object's static/dynamic status is initialized from the agent's prediction, which determines

whether its world-space pose is shared across frames. Users can correct an erroneous prediction through the annotation interface based on the video.

3.3 Point-Guided Optimization

Point-guided optimization repairs the automatic reconstruction in two stages (Alg. 1). The HOI keypoint solver aligns object poses through least-squares matching and adjusts human limbs using inverse kinematics (IK). The HOI optimizer then jointly refines human and object poses through gradient-based optimization to improve contact and physical consistency.

3.3.1 HOI Keypoint Solver

To achieve faster optimization while ensuring accuracy, we use point matching to align keypoint pairs and optimize the object pose. We then refine the human limb positions using IK.

Our core algorithm uses least squares to solve two point-alignment problems: 3D–3D spatial alignment and 3D–2D projection alignment, as shown in Eq. 1, where the optimization targets are $\mathbf{R}_o$ and $\mathbf{t}_o$.

$$\begin{aligned} \mathbf{r}_i^{3D}(\boldsymbol{\theta}) &= \sqrt{w_{3D}}\left(\mathbf{R}_o\,\mathbf{p}_i^{3D} + \mathbf{t}_o - \mathbf{q}_i^{3D}\right), \\ \mathbf{r}_j^{2D}(\boldsymbol{\theta}) &= \sqrt{w_{2D}}\left(\pi_{\mathbf{K}}\left(\mathbf{R}_o\,\mathbf{p}_j^{2D} + \mathbf{t}_o\right) - \mathbf{q}_j^{2D}\right), \\ \boldsymbol{\theta}^* &= \arg\min_{\boldsymbol{\theta}} \frac{1}{2}\left(\sum_{i=1}^{N}\left\|\mathbf{r}_i^{3D}(\boldsymbol{\theta})\right\|_2^2 + \sum_{j=1}^{M}\left\|\mathbf{r}_j^{2D}(\boldsymbol{\theta})\right\|_2^2\right). \end{aligned} \tag{1}$$

By aligning the annotated point pairs, we can track the object's pose and spatial position, as well as achieve coarse alignment of the contact regions. Since the limbs are the most critical parts for interaction and are prone to spatial misalignment, we apply a separate IK-based quick optimization specifically for the limbs.

3.3.2 HOI Optimizer

Building on the initialization provided by the HOI solver, we propose an HOI optimizer to further refine the physical properties of Human–Object Interactions by optimizing θ_h, $\mathbf{R}_o$, and $\mathbf{t}_o$. We optimize object and local body/hand poses while keeping human shape and global root motion fixed. Explicit contact pairs replace automatic contact constraints and remain active throughout refinement. The objective combines visual alignment, motion preservation, contact, and physical consistency:

$$E_{\text{total}} = E_{\text{visual}} + E_{\text{prior}} + E_{\text{temporal}} + E_{\text{contact}} + E_{\text{collision}} + E_{\text{support}}. \tag{2}$$

For an object identified as ground-supported, E_{support} penalizes the distance between its predicted support region and the estimated ground plane, while $E_{\text{collision}}$ prevents penetration. The support term is set to zero when the interaction does not require ground support.

Implementation details and the optimization schedule are provided in Sec. B.4.

4 Physics-Based HOI Motion Refinement

Our reconstruction pipeline provides not only temporally coherent human and object motion, but also fine-grained human–object contact annotations that specify when contact occurs, which human body region is involved, and where the contact lies on the object surface. Building on the generic motion-tracking objectives, control regularizers, and PPO training procedure of ProtoMotions 3 Tessler et al. (2025), we augment a standard physics-based character control policy with a fine-grained human–object contact reward that encourages simulated contacts to agree with these annotations. Through this refinement, we aim to assess whether the reconstructed sequences provide usable references for simulated character control and to improve human–object contact quality by reducing interpenetration and unintended gaps at annotated contacts.

Algorithm 1 Point-Guided Optimization

Require: Video frames with annotated 3D-3D and 3D-2D correspondences; iteration budget N_{iter}

Initialize human and object poses from automatic reconstruction
for each annotated frame t **do**
    Estimate object pose $(\mathbf{R}_t, \mathbf{t}_t)$ by weighted least squares
end for
Fuse valid object poses with the tracked trajectory
Run IK to refine human body pose parameters based on 3D keypoints
Fix human shape and root motion; use annotated contact constraints
for $i = 1$ to N_{iter} **do**
    Sample frames and optimize local human and object poses with visual, contact, collision, support, and motion losses using **Adam**
end for
return Optimized human and object poses

To focus on these goals, we prescribe the reconstructed trajectories of dynamic objects kinematically and disable their gravity, while retaining human–object collisions. This design decouples character refinement from open-ended manipulation control over objects with widely varying geometry and scale, and focuses learning on physically plausible character motion under prescribed object motion and improved human–object contact quality.

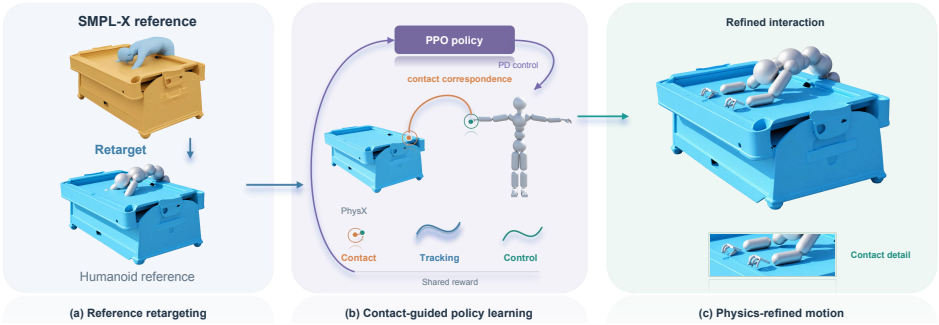


Figure 4: Physics-based HOI motion refinement. (a) Our reconstruction provides SMPL-X and object motion together with fine-grained human–object contact annotations. (b) A standard physics-based character control policy is augmented with our fine-grained human–object contact reward and trained using one shared configuration across all sequences. (c) The resulting physically simulated interaction and a close-up of its contact region.

4.1 Fine-Grained Human–Object Contact Reward

The simulated humanoid consists of articulated rigid links, and the simulator reports contact points between these links and the object. Each annotated contact group g specifies a set of eligible body links $\mathcal{B}_g$, an object-local contact target $\mathbf{x}^O_{t,g}$, and the temporal interval in which the contact is active.

At time t, we transform the object-local contact target into world coordinates using the current object pose. We then measure its distance to the nearest simulated contact point involving an eligible body link:

$$\begin{aligned} \mathbf{y}_{t,g} &= \mathbf{R}^O_t \mathbf{x}^O_{t,g} + \mathbf{t}^O_t, \\ d^2_{t,g} &= \min_{\mathbf{c} \in \mathcal{C}_{t,g}} \|\mathbf{c} - \mathbf{y}_{t,g}\|_2^2, \end{aligned} \tag{3}$$

where $\mathcal{C}_{t,g}$ contains the simulated contact points between the object and links in $\mathcal{B}_g$. We set $d^2_{t,g} = +\infty$ when no eligible contact point exists.

For the annotated contact groups $\mathcal{G}_t$ active at time t, the fine-grained human–object contact reward is

$$r^{\text{HOI}}_t = \frac{1}{|\mathcal{G}_t|} \sum_{g \in \mathcal{G}_t} \exp\left(-\frac{d^2_{t,g}}{\sigma^2} \right), \tag{4}$$

Table 1: Automatic reconstruction on InterCAP. Baseline scores are from CARI4D Xie et al. (2026); all five metrics follow its protocol (cm).

Method	CD-h ↓	CD-o ↓	CD-c ↓	Acc-h ↓	Acc-o ↓
InterTrack Xie et al. (2025)	34.79	40.37	33.53	4.34	5.26
VisTracker Xie et al. (2023)	16.12	27.41	20.17	0.98	1.25
CARI4D	11.06	**15.69**	12.88	1.25	0.82
Ours w/o contact	10.22	16.67	12.59	**0.79**	0.77
Ours	**10.21**	16.57	**12.53**	0.84	**0.76**

Table 2: Point-pair repair on 200 wild sequences. Values are medians of sequence-level scores; Support denotes the ground support gap and uses only the ground-supported subset.

Method	Contact ↓ (mm)	Pen. ↓ (mm)	Jerk ↓ (m/s^3)	Support ↓ (mm)
CARI4D	89.63	20.14	67.67	246.47
Ours (automatic)	74.15	33.99	**33.16**	86.41
Ours + point repair	**17.15**	**16.04**	39.05	**42.69**

and is zero when $\mathcal{G}_t$ is empty. The reward is high only when the specified body region establishes a simulated contact near the annotated contact target. The same formulation applies to different objects, body regions, and interaction phases without action-specific reward engineering.

4.2 Physics-Based Motion Tracking

We combine the standard motion-tracking rewards and control regularizers from ProtoMotions 3 into r_t^{base} and augment them with our fine-grained human–object contact reward:

$$r_t = r_t^{\text{base}} + w_c r_t^{\text{HOI}}, \tag{5}$$

where w_c controls the strength of the fine-grained human–object contact reward. The motion-tracking objective preserves the reconstructed reference motion, while r_t^{HOI} encourages the policy to realize the contact annotations as simulated contacts.

Using a single training configuration shared across the dataset, we train a separate physics-based character control policy for each sequence. The policy is optimized with PPO Schulman et al. (2017). Detailed simulation, reward, and training settings are provided in Sec. C.

5 Experiments

5.1 Experimental Setup

Data. Following CARI4D Xie et al. (2026), we use the same selected InterCAP Huang et al. (2024) test videos and camera views to evaluate automatic reconstruction against 3D ground truth. However, InterCAP is captured in controlled indoor environments and contains relatively simple interactions, providing limited discrimination on challenging cases. We therefore construct a complementary in-the-wild benchmark of 200 sequences collected from internet videos and smartphone recordings, covering 92 object categories. All sequences have our point annotations, enabling evaluation of repair and physical refinement under diverse in-the-wild conditions.

Metrics and protocol. On InterCAP, we follow CARI4D and report human, object, and combined-mesh Chamfer distance (CD-h, CD-o, CD-c), together with human and object acceleration error (Acc-h, Acc-o). On the wild benchmark, we evaluate four aspects of physical consistency: contact distance, penetration, motion smoothness of the object, and, for ground-supported objects, support gap. Detailed evaluation protocols are provided in Sec. B.

5.2 Agent-Based Automatic Reconstruction

We use Qwen3-VL-8B-Instruct Bai et al. (2025) as the vision–language agent in our automatic reconstruction pipeline. We compare against template-free InterTrack Xie et al. (2025), visibility-aware VisTracker Xie et al. (2023), and category-agnostic CARI4D Xie et al. (2026). Tab. 1 reports the quantitative comparison on InterCAP, including a variant without contact guidance. Score differences are modest on this relatively simple indoor benchmark, whose global Chamfer and acceleration metrics may underrepresent local contact improvements. Fig. 7 shows more visible gains in human–object alignment and contact in the illustrated cases.

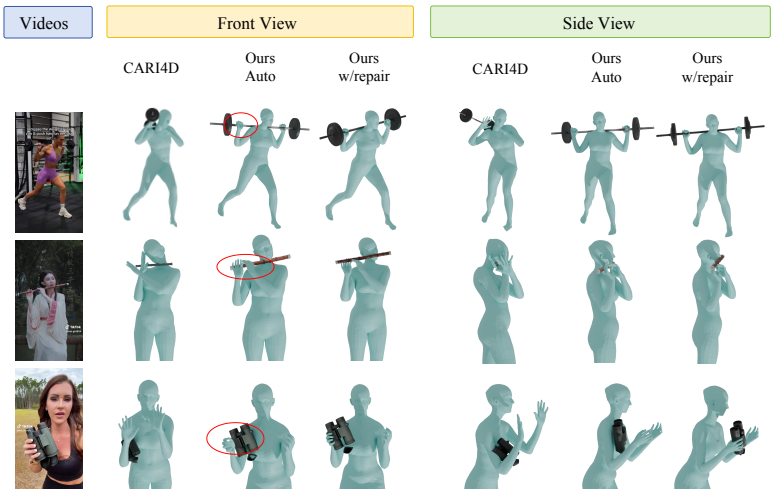


Figure 5: Qualitative comparison on the wild benchmark. Input video frames are shown on the left; front and side views are arranged for CARI4D, our agent-based automatic reconstruction (Ours Auto), and our point-repaired reconstruction (Ours w/repair).

5.3 Point-Pair Repair

Tab. 2 compares CARI4D, our automatic reconstruction, and our automatic reconstruction with point-pair repair. The before/after comparison uses the same initialization and object mesh, isolating the effect of sparse manual constraints. Contact distance measures satisfaction of the annotated contacts, not independent reconstruction accuracy. We therefore report it alongside penetration, object-motion smoothness, and ground support. Point repair improves contact consistency, reduces penetration and support gap, and maintains smooth motion.

The point pairs target errors in relative human–object depth, object pose, and interacting-limb placement. Contact pairs constrain relative 3D placement, while object-to-image correspondences correct pose errors visible in the input video; temporal fusion propagates these corrections between annotated frames. Fig. 5 shows the visible improvements in object alignment and contact quality from the front and side views.

5.4 Annotation Efficiency and Reconstruction Quality

We compare three ways to repair automatic reconstructions on 20 in-the-wild sequences selected for visible errors in the automatic results: Blender-like object-pose adjustment, painted contact-region annotation, and our sparse point-pair annotation. For the first two workflows, contact regions derived from pose edits or painted directly provide constraints for the same HOI optimization. Two professional annotators independently complete all three workflows from the same initialization.

Tab. 3 reports mean total manual operation time, penetration depth, human ratings, and failure rate. Five independent reviewers assessed rendered videos in randomized order with method names hidden. Results are averaged over the two annotators and 20 sequences; workflow and review details are provided in Sec. D.

Our point pairs require less manual time while achieving higher quality (Tab. 3). As illustrated in Fig. 6, Blender-like pose edits are difficult to propagate between annotated frames; the inferred contact regions used for propagation can be inaccurate. Even carefully painted contact regions remain spatially coarse and underconstrain the relative pose compared with explicit point pairs.

5.5 Physical Refinement

Tab. 4 compares the input reference, our policy, and RePHO on the fixed 45-motion test subset. We measure motion error relative to the reference, joint-angle jerk, and penetration on the visual object mesh. We adapt RePHO Huang et al. (2026) to use the same prescribed dynamic-object trajectories as our method during both training and evaluation.

Metric	Blender	Painted	Ours
Time (s) ↓	54.47	86.74	**40.63**
Pen. (mm) ↓	33.58	30.34	**10.87**
Human rating ↑	67.6	70.2	**89.8**
Failure ↓	7/20	5/20	**0/20**

Table 3: Annotation efficiency and reconstruction quality on 20 in-the-wild sequences. Time is mean total manual operation time; Pen. is penetration-depth P95; Human rating averages five independent reviewers' scores on a 0–100 scale from rendered videos shown in random order without method names. Failure is the fraction judged to have obvious object tracking failure, such as severe interpenetration or a physically implausible pose.

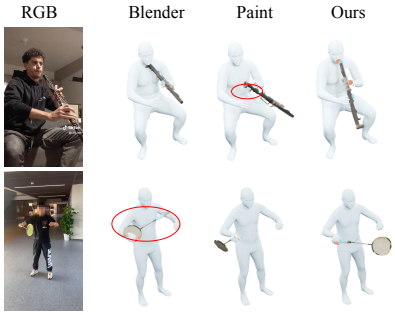


Figure 6: Qualitative comparison of the annotation workflows.

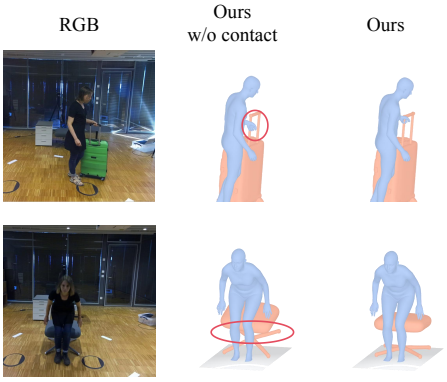


Figure 7: InterCAP reconstruction with and without contact guidance.

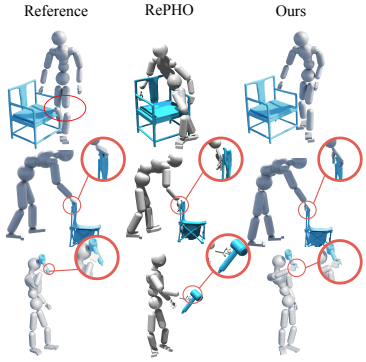


Figure 8: Reference and physically refined HOI motions.

Table 4: Quantitative comparison on the fixed 45-motion test subset. Values are per-motion macro averages.

Method	MPJPE (mm) ↓	Angular jerk ($\mathrm{rad/s^3}$) ↓	Pen. (mm) ↓	Pen. frames (%) ↓
Input reference	–	366.96	5.24	38.56
RePHO Huang et al. (2026)	153.46	981.15	3.95	28.43
Ours	**94.62**	**183.38**	**2.79**	**17.96**

MPJPE denotes full-joint global position error without pelvis alignment. Angular jerk is computed from local joint angles. Pen. is the mean human–object penetration depth, including zero-penetration frames; Pen. frames is the fraction of frames exceeding 5 mm.

Fig. 8 compares reference and refined motion across interactions, illustrating contact quality and motion preservation.

6 Conclusion

4DHOISolver combines agent-based reconstruction, sparse point-pair repair, and physical refinement to recover 4D HOI motion from in-the-wild monocular videos. This progression efficiently corrects difficult pose and contact errors while improving physical plausibility across diverse interactions. Under severe occlusion, the method remains dependent on the quality of the initial reconstruction, and currently supports only rigid objects.

AI USE STATEMENT

We use Qwen3-VL-8B-Instruct to assist automatic HOI reconstruction, as described in the method and experiments. Generative AI tools also assisted the development of our annotation tool and its web frontend, and helped polish the manuscript's language. The researchers reviewed and verified the experimental code, including AI-assisted components. The authors take responsibility for the final manuscript, implementation, and reported results.

ETHICS STATEMENT

Our evaluation benchmark includes internet videos collected through TikTok's built-in download feature and smartphone recordings collected by the authors. For TikTok videos, our release will provide source video URLs rather than redistribute the video files. Users must obtain the videos directly from the platform, subject to its applicable terms and creators' download settings. Participants in our smartphone recordings consented to the public release of their videos.

REPRODUCIBILITY STATEMENT

The automatic reconstruction pipeline, annotation interface, and point-guided optimization are described in Secs. 3.1, 3.2, and 3.3, respectively. Section 4 specifies the physical refinement objective and policy training procedure, with simulation and training settings in Sec. C. The experimental setup and evaluation protocols are described in Sec. 5.1, including the RGB-only input used for both InterCAP and in-the-wild evaluation. These descriptions provide the basis for reproducing the reconstruction, repair, and physical refinement stages.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15935–15946, jun 2022.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Yanjun Wang, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, Chen Change Loy, Lei Yang, and Ziwei Liu. SMPLer-X: Scaling up expressive human pose and shape estimation. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xingyu Chen, Zhuheng Song, Xiaoke Jiang, Yaoqing Hu, Junzhi Yu, and Lei Zhang. HandOS: 3D hand reconstruction in one stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17304–17314, June 2025.

Alpár Cseke, Shashank Tripathi, Sai Kumar Dwivedi, Arjun S. Lakshmipathy, Agniv Chatterjee, Michael J. Black, and Dimitrios Tzionas. PICO: Reconstructing 3D people in contact with objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1783–1794, June 2025.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi.

Objaverse-XL: A universe of 10M+ 3D objects. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35799–35813, 2023.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22605–22615, June 2025.

Pei Geng, Shanshan Zhang, and Jian Yang. CrossHOI: Learning cross-view representations for monocular 3D human-object interaction reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7121–7130, June 2026.

Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14783–14794, October 2023.

Dingbang Huang, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Recovering physically plausible human-object interactions from monocular videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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*, pp. 1–16, 2024.

Chaofan Huo, Ye Shi, and Jingya Wang. Monocular human-object reconstruction in the wild. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5547–5555, 2024.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1737–1747, June 2024.

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

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is better to track together. In *Computer Vision – ECCV 2024*, pp. 18–35, Cham, 2025. Springer Nature Switzerland.

Ao Li, Jinpeng Liu, Yixuan Zhu, and Yansong Tang. ScoreHOI: Physically plausible reconstruction of human-object interaction via score-guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7592–7602, October 2025a.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6), 2023. doi: 10.1145/3618333.

Jiaxin Li, Yuxiang Wu, Zhenkai Zhang, Xinrui Shi, Haoyuan Wang, Yichen Zhao, Linxiang Su, Chenyang Yu, Mingyu Zhang, Yifan Ding, Boran Wen, Li Zhang, Ruiyang Liu, and Yong-Lu Li. HAT-4D: Lifting monocular video for 4D multi-object interactions via human-agent collaboration, 2026.

Jiefeng Li, Jinkun Cao, Haotian Zhang, Davis Rempe, Jan Kautz, Umar Iqbal, and Ye Yuan. GENMO: A GENeralist model for human MOtion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11766–11776, October 2025b.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3D whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21159–21168, June 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023b.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1769–1782, June 2025.

Jiaxin Lu, Chun-Hao Paul Huang, Uttaran Bhattacharya, Qixing Huang, and Yi Zhou. HUMOTO: A 4D dataset of Mocap human object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10886–10897, October 2025.

Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10895–10904, October 2023.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, June 2024.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics*, 37(4):143:1–143:14, 2018. doi: 10.1145/3197517.3201311.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, June 2025.

Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5099–5108, 2017.

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. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Proceedings*, 2024.

Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023.

Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J. Black. WHAM: Reconstructing world-grounded humans with accurate 3D motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2070–2080, June 2024.

Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-one-stage expressive human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1834–1843, June 2024.

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. Sam 3d: 3dfy anything in images, 2025. URL https://arxiv.org/abs/2511.16624.

Chen Tessler, Yifeng Jiang, Xue Bin Peng, Erwin Coumans, Yi Shi, Haotian Zhang, Davis Rempe, Gal Chechik, and Sanja Fidler. Protomotions3: An open-source framework for humanoid simulation and control. https://github.com/NVLabs/ProtoMotions/, 2025.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-based imitation of dynamic human-object interaction, 2023.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. SkillMimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17540–17549, June 2025.

Yufu Wang, Ziyun Wang, Lingjie Liu, and Kostas Daniilidis. TRAM: Global trajectory and motion of 3D humans from in-the-wild videos. In *Computer Vision – ECCV 2024*, volume 15069, pp. 467–487, 2024.

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. Arthoi: Taming foundation models for monocular 4d reconstruction of hand-articulated-object interactions, 2026. URL https://arxiv.org/abs/2603.25791.

Boran Wen, Dingbang Huang, Zichen Zhang, Jiahong Zhou, Jianbin Deng, Jingyu Gong, Yulong Chen, Lizhuang Ma, and Yong-Lu Li. Reconstructing in-the-wild open-vocabulary human-object interactions, June 2025.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, June 2024.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. *Tech report*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3D latents for scalable and versatile 3D generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21469–21480, June 2025b.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single RGB camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4757–4768, June 2023.

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

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2026.

Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. InstantMesh: Efficient 3D mesh generation from a single image with sparse-view large reconstruction models, 2024.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12266–12277, June 2025.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. Interprior: Scaling generative control for physics-based human-object interactions, 2026. URL https://arxiv.org/abs/2602.06035.

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, 2026. URL https://arxiv.org/abs/2605.17014.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, and Zheng-Jun Zha. LEMON: Learning 3D human-object interaction relation from 2D images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16284–16295, June 2024.

Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21222–21232, June 2023.

Qiao Yu, Xianzhi Li, Yuan Tang, Xu Han, Long Hu, Yixue Hao, and Min Chen. Fancy123: One image to high-quality 3D mesh generation via plug-and-play deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 595–604, June 2025a.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. 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)*, pp. 27716–27726, June 2025b.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. SimPoE: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7159–7169, June 2021.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8834–8845, June 2023.

Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. HOI-M^3: Capture multiple humans and objects interaction within contextual environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 516–526, June 2024.

Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular RGB image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2354–2364, October 2019.

APPENDIX

A DETAILS OF 4DHOISOLVER

This section provides additional details on our automatic reconstruction and point-guided repair pipeline, covering agent-based reconstruction, human contact-point definitions, an optional contact-proposal model, and the annotation interface. Fig. 9 shows how these components fit together.

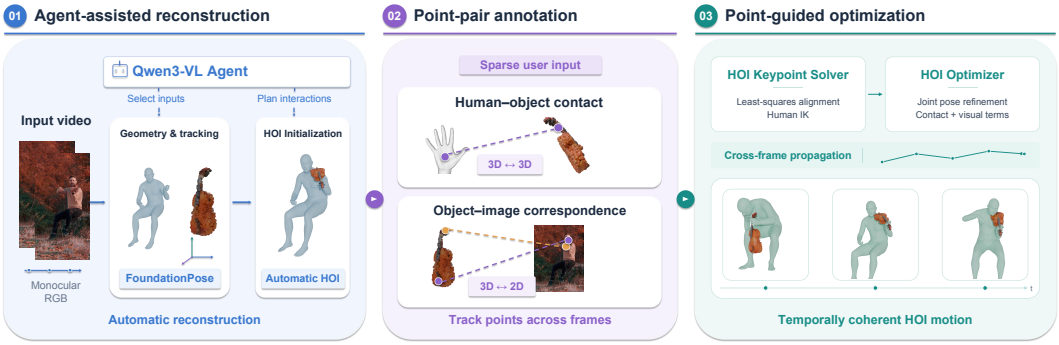


Figure 9: Overview of automatic reconstruction and point-guided repair. Given a monocular RGB video, a vision–language agent assists input selection and interaction planning, while geometry and tracking modules initialize the human–object motion. When correction is needed, users annotate sparse 3D human–object contact pairs and 3D-to-2D object–image correspondences, which are propagated across frames. Our point-guided solver integrates these constraints across the sequence to correct the reconstruction and recover temporally coherent HOI motion.

A.1 AUTOMATIC RECONSTRUCTION DETAILS

We provide additional details for the automatic reconstruction pipeline described in Sec. 3.1. The pipeline produces a temporally complete human motion sequence, an object mesh, and an object pose trajectory, which serve as the initialization for agent-assisted optimization.

A.1.1 INTERACTING INSTANCE AND MESH SELECTION

We sample frames throughout the video and detect candidate people and objects. A vision-language agent identifies the interacting pair and maintains the identity of the target object across sampled frames. We then propagate the selected masks in both temporal directions with SAM2. To obtain a reliable object mesh, we rank candidate frames by mask completeness, visible area, and image quality; the agent reviews the strongest candidates before a selected frame is passed to SAM3D. A single object scale is estimated from the calibrated monocular depth and rendered mesh, and is kept fixed throughout the sequence.

A.1.2 HUMAN AND OBJECT POSE INITIALIZATION

A global human-motion estimator provides body motion, while a hand estimator supplies detailed hand poses; these predictions are fused into a SMPL-X sequence. Given the object mesh, RGB frames, masks, and predicted depth, FoundationPose initializes and tracks the object's orientation and translation. We validate FoundationPose estimates against the observed object mask and calibrated scene depth using rendered silhouette and depth, and reject large temporal jumps. Failed frames trigger bounded re-registration and neighbor-seeded tracking. Remaining missing poses are completed from nearby estimates, while temporal pose spikes are smoothed. The resulting trajectory undergoes depth-guided pose refinement before joint HOI optimization.

A.1.3 DEPTH-BASED ALIGNMENT AND TEMPORAL STABILIZATION

We predict scene depth from the monocular RGB frames and calibrate it against the rendered human depth. No registered sensor depth or ground-truth 3D geometry is used as reconstruction input. We

first align the human root translation to the calibrated lower-body depth while retaining the estimated articulation and shape. We then refine the object's translation against its visible predicted depth points, keeping its FoundationPose orientation and estimated scale fixed. A robust surface-distance objective limits the influence of occlusion and depth prediction errors. Finally, we smooth the object translation trajectory with a second-order temporal penalty, assigning less weight to frames with unreliable depth agreement. This separates the complementary roles of the two sources: FoundationPose supplies object orientation, whereas calibrated monocular depth anchors its translation.

A.1.4 Agent Input and Semantic Priors

The refinement agent receives temporally sampled full-frame images together with object-centered crops. The images display the human and object masks in distinct colors and include frame indices; the prompt also specifies the object category and approximate size. The agent predicts contact intervals and relevant body regions, and determines whether the object should remain static or move with the interaction. For a static object, it additionally identifies whether ground support is expected. These outputs are semantic priors rather than direct 3D pose predictions.

A.1.5 Prior-Guided HOI Refinement

The predicted contact intervals activate a surface-proximity objective for the relevant human joints or kinematic chains. This lets the body approach the reconstructed object during interaction without forcing unrelated body regions to move. Scene priors control the object optimization: static objects are encouraged to maintain a consistent world pose and, when appropriate, are placed on the estimated ground through their support geometry; dynamic objects are not subjected to a ground-contact constraint. The refinement retains the initialization as a reference and combines these semantic objectives with geometric and motion-consistency terms. Thus, the agent specifies *what* physical relationship to enforce, while the optimizer determines the corresponding 3D adjustment.

A.2 Human Keypoint Definition

In Sec. 3.2, we mention that, for convenient and accurate interaction annotation, we predefine keypoints on the human body to serve as annotation targets. We first subdivide the human body parts by evenly splitting each major joint into front, back, left, and right regions. Then, for each subdivided part, we select a central point as the interaction keypoint. For the hand interaction points, we selected corresponding locations on all five fingers, as well as points on both the palm and the back of the hand.

In total, we defined 87 keypoints. Fig. 10 visualizes our human body keypoints, and Tab.5 lists the names of the joints.

A.3 InterPoint Contact Proposals

InterPoint is an optional assistant in our annotation app that reduces manual point selection by proposing human–object contact pairs. Given a video frame and the reconstructed object point cloud, it predicts which of the 87 predefined human keypoints are in contact and their corresponding object-surface points. Annotators inspect and correct these proposals in the app; only the accepted pairs are used by the point-guided solver.

InterPoint combines image semantics from LLaVA Liu et al. (2023a) with object geometry encoded by PointNet++ Qi et al. (2017). Learnable features on the shared human template represent the human keypoints. Semantic features condition the human and object features through feature-wise modulation, and an interaction transformer combines them. A human head predicts contact probabilities, while an object head scores sampled object points for each human keypoint. At inference, the highest-scoring object point is proposed for each predicted human contact (Fig. 11).

We train the human head with binary cross-entropy and the object head with bidirectional contrastive matching against neighborhoods of annotated object contact points. The combined objective is

$$\mathcal{L}_{\text{InterPoint}} = \lambda_h \mathcal{L}_{\text{human}} + \lambda_{h\to o}\mathcal{L}_{h\to o} + \lambda_{o\to h}\mathcal{L}_{o\to h}. \tag{6}$$

This neighborhood supervision accommodates small differences between a predicted surface point and the annotated contact location. Training uses 1,024 sampled object points and the 87-keypoint

Table 5: Human joint-tree.

Main-Joints	Sub-Joints
leftForeArm	back, pinky, wrist, thumb
rightForeArm	back, pinky, wrist, thumb
leftUpperArm	up, down, back, front
rightUpperArm	up, down, back, front
leftShoulder	front, back
rightShoulder	front, back
leftHand	back, palm, Thumb, Index, Middle, Ring, Pinky
rightHand	back, palm, Thumb, Index, Middle, Ring, Pinky
leftUpperLeg	inner, outer, front, back
rightUpperLeg	inner, outer, front, back
leftLowerLeg	front, outer, back, inner
rightLowerLeg	front, outer, back, inner
leftFoot	ToeBase, instep, sole
rightFoot	ToeBase, instep, sole
upperSpine	back, right, front, left
middleSpine	front, right, back, left
leftNeck	front, back
rightNeck	back, front
hip	front, left, right, back
buttocks_left	buttocks_left
buttocks_right	buttocks_right
head	mouth, chin, headtop, cheek_left, cheek_right
rightElbow	back, front
leftElbow	back, front
leftKnee	back, front
rightKnee	back, front

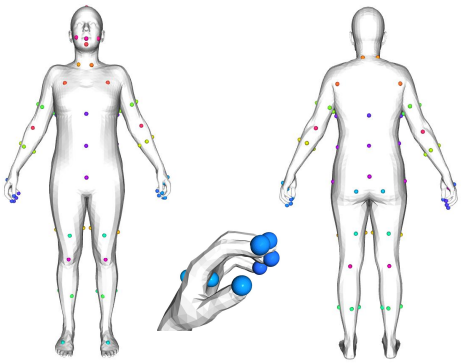


Figure 10: Human keypoint definition.

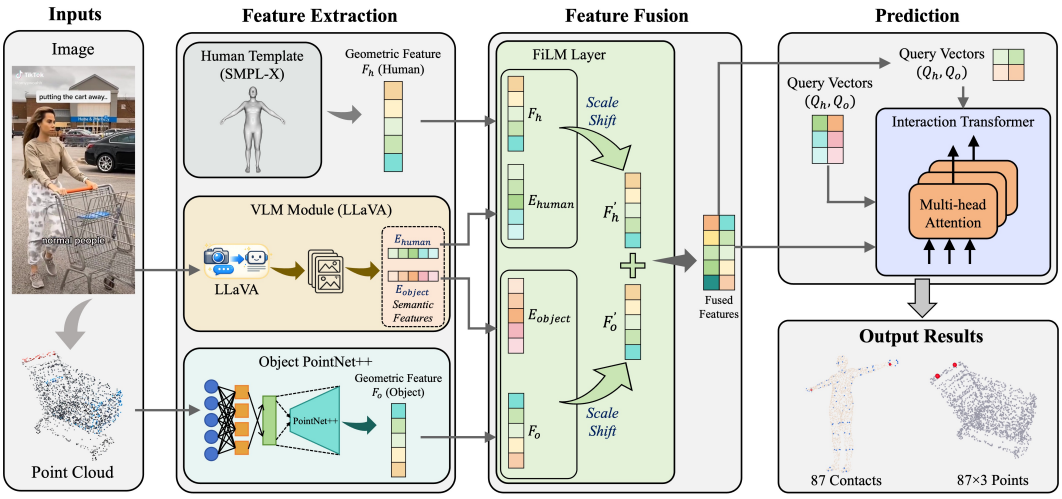


Figure 11: InterPoint architecture. Image semantics and object point-cloud features condition human and object representations, which an interaction transformer uses to predict human contact keypoints and their corresponding object surface points.

human definition above. We optimize for 80 epochs with AdamW, a batch size of 16, and a learning rate of 3×10^{-5}.

A.4 Annotation App

To efficiently repair HOI reconstructions, we develop a custom interactive annotation application. Fig. 12 displays the user interface, and Fig. 13 illustrates the step-by-step workflow. The workflow includes optional InterPoint proposals; users verify or edit these candidates before optimization.

User Interface & Toolbar. Our system provides an intuitive toolbar for rapid interaction control:

- **Video Timeline:** Users can drag the progress bar or play the video to locate specific frames that require annotation.

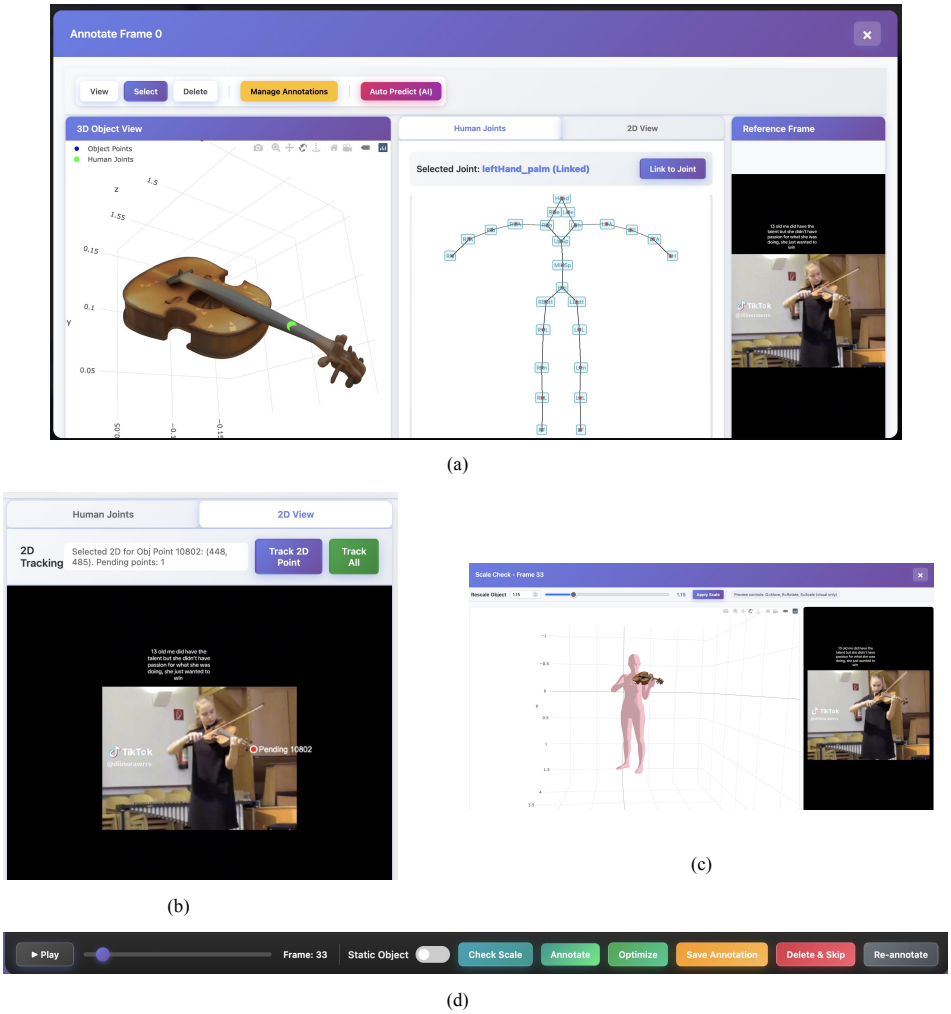


Figure 12: **Overview of our interactive annotation system.** (a) The interface for selecting object interaction points and assigning corresponding human joints. (b) The 2D point tracking interface for verifying temporal correspondences. (c) The viewer for inspecting the fast optimization and 4D reconstruction results. (d) The interactive toolbar for system controls.

- **Check Scale:** This function allows annotators to inspect and adjust the automatically reconstructed object scale, as well as visually verify the optimization results.
- **Static Object:** A toggle designed for stationary items. Users can correct the agent-predicted motion attribute through this control.
- **Optimize:** Triggers our HOI keypoint solver for rapid geometric alignment.

Annotation Workflow. As shown in the workflow diagram, a complete annotation cycle proceeds as follows:

- **InterPoint Auto-prediction:** Annotators can request candidate contact pairs from InterPoint and inspect them before accepting or editing them.
- **3D Point & Human Joint Selection:** Users can correct a reconstruction by selecting 3D contact points on the object and assigning them to corresponding human joints via a hierarchical joint-tree menu.
- **2D Point Tracking:** Users specify the 2D image projections of the selected 3D object points. The app then tracks these points temporally, allowing annotators to easily correct any tracking drift.

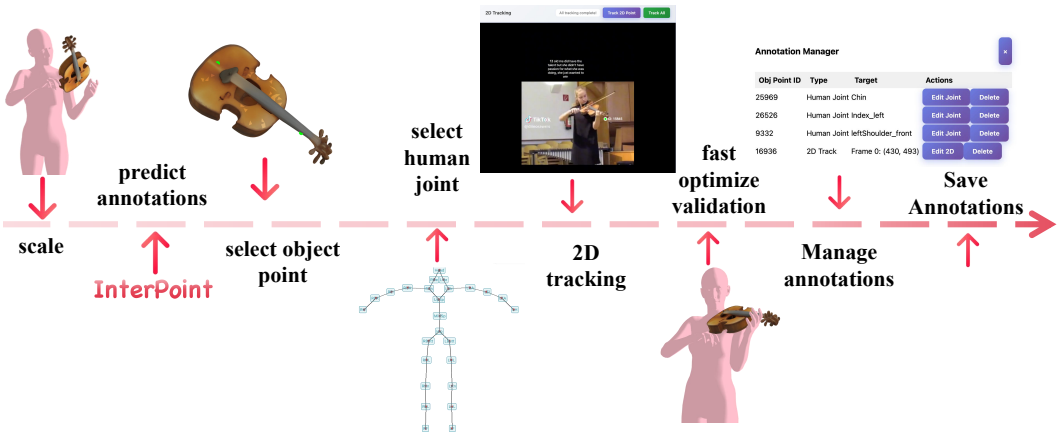


Figure 13: **Workflow of our human-in-the-loop annotation pipeline.** The timeline illustrates the step-by-step process from initial video preprocessing and interactive point annotation to the final 4D HOI reconstruction.

- **Manage Annotations:** Annotators can flexibly edit, re-link, or delete specific contact pairs or tracking points without discarding the valid parts, greatly improving efficiency.
- **Fast Optimization & Verification:** Finally, users utilize the *Optimize* function to quickly solve the 4D HOI reconstruction. This allows them to seamlessly verify the physical plausibility and temporal consistency of their current annotations and make further adjustments if necessary.

B Evaluation Protocol

B.1 Evaluation Sets

Automatic reconstruction is evaluated on the same selected InterCAP Huang et al. (2024) test videos and camera views as CARI4D Xie et al. (2026), for which 3D ground truth is available for evaluation only; the reconstruction input is monocular RGB and does not include sensor depth. Point-pair repair is evaluated on a separate in-the-wild benchmark of 200 sequences spanning 92 object categories. These sequences come from internet videos and consented smartphone recordings, and each has human–object contact pairs and object-to-image correspondences. Physical refinement uses a subset of this benchmark, described below.

B.2 Automatic Reconstruction Evaluation

For the supplementary InterCAP experiment, we sample 6,000 points per mesh and compute symmetric Chamfer distances for the human, object, and their combined surfaces (CD-h, CD-o, and CD-c). Acceleration errors use second-order frame differences of 22 human joints and the object center, without frame-rate normalization. Both Chamfer distances and acceleration errors are reported in centimeters. A first-frame Sim(3) alignment is fitted using only the human mesh and then applied to both human and object trajectories; each variant is aligned independently. We average errors over frames and then sequences.

B.3 Point-Repair Evaluation

All methods use the same frames and contact annotations. For annotated human contact point $h_{t,k}$ and object surface O_t, we report $\mathrm{Contact} = \mathrm{median}_{t,k} \min_{y \in O_t} \|h_{t,k} - y\|$. Let $d_{t,i} > 0$ denote the inward signed distance of an object surface sample inside the human mesh; $\mathrm{Pen.} = Q_{0.95}(\{d_{t,i} : d_{t,i} > 0\})$, or zero if no samples penetrate. For object-center positions x_t at frame rate f, $\mathrm{Jerk} = \mathrm{mean}_t \|\Delta^3 x_t\| f^3$. For static objects, Support is the absolute median signed distance from the lowest 2% of object samples to a ground plane estimated from stable, low human vertices. We report the median sequence score; distances are in millimeters and Jerk is in $\mathrm{m/s^3}$.

B.4 Point-Guided Optimization Implementation

For Eq. 1, object poses and human contact targets are expressed in camera coordinates. Levenberg–Marquardt solves each constrained frame, warm-started from the preceding valid solution. Residual multipliers are 900 for metric 3D coordinates and 2 for pixel coordinates, corresponding to $w_{3D} = 900^2$ and $w_{2D} = 2^2$. Failed or underconstrained solutions are excluded from the observations used for fusion.

The solved and tracked poses are transformed into a shared world frame. Confidence-weighted observations and second- and third-order temporal differences regularize the fused trajectory. Translation and sign-consistent quaternion coordinates are smoothed separately, followed by quaternion normalization and restoration of reliable least-squares anchors. IK adjusts the interacting human limbs to the repaired object contacts before joint refinement. The IK-adjusted human pose and repaired object trajectory initialize the subsequent local residual optimization.

The point-guided implementation uses LM pose fitting followed by world-space trajectory fusion. It retains all numerically valid LS solutions in the hard-LS configuration, regularizes the pose observations, and restores reliable anchors after fusion. Human shape and root motion remain frozen; local body/hand residuals are bounded. Dynamic-object updates are limited to 3 cm translation and 10 degrees rotation around the repaired seed. Static ground-supported objects use a shared scale bounded by a factor of 1.25; final rigid ground correction is limited to 25 cm and 40 degrees.

The 1500-step interaction schedule comprises object-depth/pose repair (0–299), contact refinement (300–649), local human refinement (650–1049), and physical finalization (1050–1499). Explicit point-pair loss remains active throughout the point-guided run and replaces the generated-contact term.

$$E_{\text{point}} = \frac{1}{|\mathcal{C}_{\mathcal{B}}|} \sum_{(t,a,b)\in\mathcal{C}_{\mathcal{B}}} \max\big(0, \|\mathbf{v}^o_{t,a} - \mathbf{v}^h_{t,b}\|_2 - \epsilon\big)^2, \tag{7}$$

Here $\mathcal{C}_{\mathcal{B}}$ contains the annotated contact pairs in a sampled frame batch; the loss is defined as zero for an empty set. The tolerance ϵ is 15 mm for full-body contacts and 5 mm for fine-hand contacts, with zero attraction inside the tolerance. This differs from the automatic full-body contact loss, which penalizes deviation from a 1 cm target distance.

C Details of Physics-Based HOI Refinement

Our implementation is built upon and extends ProtoMotions 3 Tessler et al. (2025). We inherit its PPO training infrastructure, motion-library interface, simulator abstraction, and humanoid tracking formulation. We introduce the HOI-specific components, including subject-specific SMPL-X proxies, reconstructed object-scene integration, semantic human–object contact objectives, collision handling, and the evaluation pipeline described below. We use the Isaac Lab backend Mittal et al. (2025) with PhysX simulation.

C.1 Physical Refinement Setup

Starting from the reconstructed HOI sequences produced by our reconstruction stage, we convert each sequence into an independent physics-based tracking task. Each input contains a temporally aligned SMPL-X human, a reconstructed object mesh and its trajectory, and semantic interaction annotations. Using a single training configuration shared across the dataset, we train a separate physics-based character controller for each sequence to track the corresponding reconstructed motion in simulation.

The simulator runs at 120 Hz and the policy acts at 30 Hz. Human–ground and human–object collisions remain active during refinement. The reference human, object, and semantic contact targets share the same world coordinate system and are sampled at the policy time step.

C.2 Human and Scene Conversion

For each sequence, we instantiate the physics-based humanoid adopted from RePHO Huang et al. (2026) using the reconstructed SMPL-X parameters. The reconstructed body shape defines the

Table 6: Reward terms used for physics-based refinement.

Reward term	Weight
Link position tracking	0.5
Link rotation tracking	0.3
Linear velocity tracking	0.1
Angular velocity tracking	0.2
Root-height tracking	0.2
Semantic object contact	0.2
Foot-contact consistency	0.1
Consecutive-action regularization	0.02
Mechanical power regularization	10^{-5}

subject-specific humanoid, while the SMPL-X pose and root trajectory provide its reference motion. We retain the collision geometry, inertia, actuators, and joint-frame conventions of the RePHO dynamics template. The resulting humanoid has 52 links and 153 controlled hinge degrees of freedom.

A proper rotation converts the declared source up-axis to Z-up. We recompute the SMPL-X translation after this rotation so that the transformed pelvis position is preserved. Object-local vertices, object poses, and semantic contact targets receive the same coordinate transformation. A shared vertical shift then places the lowest SMPL-X surface vertex over the full sequence at ground height.

The proxy root state and joint coordinates initialize the simulator. We replace the proxy Cartesian joint positions with the 52 name-matched joints recovered directly from SMPL-X. Linear velocities are estimated by forward differences over one, two, and three frames, selecting the minimum-magnitude valid estimate for each joint and time step to suppress isolated reconstruction spikes. Boundary frames reuse the closest valid estimate. The reference rotations, angular velocities, and joint coordinates follow the converted proxy motion. Mesh cleanup and simplification remap annotated contact vertices to the nearest retained geometry and record the remapping distance.

The humanoid proxy is used for simulation and control, whereas the reconstructed SMPL-X surface is recovered for geometric evaluation. In particular, the human–object penetration metrics below are evaluated on the SMPL-X surface rather than on the coarser rigid-body collision shapes.

C.3 Policy Learning

Unless otherwise specified, we retain the PPO implementation and actor–critic architecture of ProtoMotions 3. Our modifications support HOI references, subject-specific humanoid assets, semantic contact objectives, object-aware simulation, and large-scale per-sequence training.

The actor and critic each receive a 2179-dimensional input: 778 dimensions for the current humanoid state, 1248 dimensions for the next-step reference, and the previous 153-dimensional raw action. There are no explicit object-pose or contact-state observations. The Gaussian policy output passes through $\tanh$ and is mapped to joint-range-based PD targets.

For all $B = 52$ links, the position and velocity tracking errors are

$$E(\mathbf{x}, \mathbf{x}^*) = \frac{1}{3B} \sum_{b=1}^{B} \|\mathbf{x}_b - \mathbf{x}_b^*\|_2^2 , \tag{8}$$

and the rotation error is the mean squared shortest quaternion angle. The tracking rewards use exponential coefficients 25, 5, 0.5, 0.1, and 100 for link position, link rotation, linear velocity, angular velocity, and root height, respectively. Tab.6 summarizes the reward weights.

Semantic contact groups are constructed from the annotated body regions and mapped to eligible humanoid links. Their targets are sampled from the nearest annotation frame and transformed by the current object pose. A reported object–body contact requires a force above 0.5 N and a finite world-space contact point. The spatial contact kernel uses $\sigma = 0.05$ m. Missing an expected contact yields no contact reward, while unannotated contacts are not penalized by this term.

Reference foot-contact labels are obtained when an ankle or toe is within 0.1 m of its sequence-wide minimum height and moves below 0.15 m/s, followed by a seven-frame moving average. The

Table 7: Policy and PPO configuration.

Setting	Value
Simulator	Isaac Lab / PhysX
Physics / control frequency	120 / 30 Hz
Links / action dimensions	52 / 153
Observation dimensions	2179
Parallel environments	Capacity-scaled across runs
Rollout length	32 steps
Minibatch size	16,384 transitions
Nominal training budget	3000 normalized epochs
Actor / critic hidden layers	6×1024 / 4×1024 ReLU
Actor / critic learning rate	2×10^{-5} / 10^{-4}
PPO clip / discount / GAE	0.2 / 0.99 / 0.95
Fixed action log standard deviation	-2.9
Training seed	0

corresponding penalty is the absolute difference between reference and simulated contact flags and is disabled for the first five control steps after reset. The consecutive-action term penalizes the change between successive PD targets, and the power term penalizes $\sum_j |\tau_j \dot{d}_j|$. Task rewards are normalized with running statistics and clipped to $[-5, 5]$ before PPO updates.

Resets start at the first frame with probability 0.2 and otherwise sample a valid reference time, synchronizing the humanoid state, scene, and contact targets. Tab.7 summarizes the policy and PPO settings.

C.4 Evaluation Metrics

Global MPJPE. For a sequence of T frames and the full set of $J = 52$ joints, we measure the world-space deviation from the reconstructed reference as

$$\mathrm{MPJPE} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \mathbf{x}_{t,j} - \mathbf{x}_{t,j}^{\mathrm{ref}} \right\|_2 . \tag{9}$$

We report this global error in millimeters without pelvis alignment.

Joint-angle jerk. Let $\mathbf{R}_{t,j}$ denote the local rotation of a non-root joint. We first compute its discrete angular velocity using the rotation logarithm,

$$\boldsymbol{\omega}_{t,j} = \frac{\mathrm{Log}(\mathbf{R}_{t,j}^{\mathsf{T}} \mathbf{R}_{t+1,j})}{\Delta t}, \tag{10}$$

and approximate angular jerk by the second finite difference of angular velocity,

$$\mathbf{j}_{t,j} = \frac{\boldsymbol{\omega}_{t+2,j} - 2\boldsymbol{\omega}_{t+1,j} + \boldsymbol{\omega}_{t,j}}{\Delta t^2}. \tag{11}$$

The reported value is the mean of $\|\mathbf{j}_{t,j}\|_2$ over time and all 51 non-root joints, in $\mathrm{rad/s^3}$.

Human–object penetration. At each frame, we transform the reconstructed SMPL-X surface vertices into the object coordinate frame and evaluate their signed distances to the visual object mesh. The framewise penetration depth is

$$p_t = \max_v \max(-\mathrm{SDF}(\mathbf{v}_{t,v}), 0) . \tag{12}$$

Pen. is the temporal mean of p_t, including frames with zero penetration. *Pen. frames* is the percentage of frames for which $p_t > 5$ mm. Reliable signed inside–outside classification requires a closed surface. We therefore skip penetration evaluation for sequences whose visual object mesh is non-watertight or for which signed distance cannot be computed robustly.

All metrics are first computed independently for each motion and then macro-averaged, preventing longer sequences from dominating the reported results.

D Annotation Efficiency and Quality Comparison

D.1 Study Setup

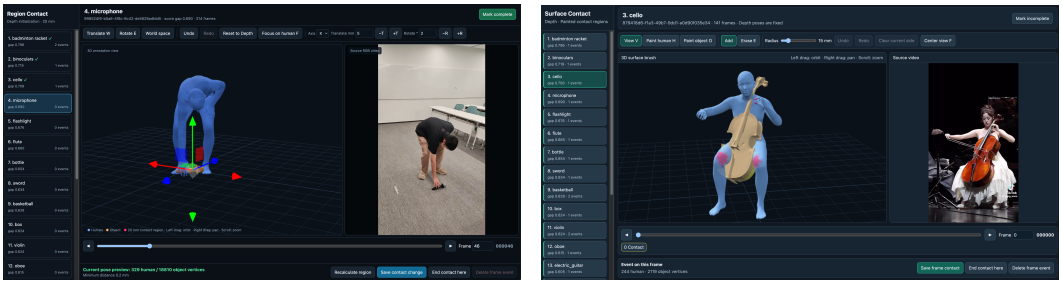
The comparison in Sec. 5.4 uses 20 in-the-wild sequences selected for visible errors in the automatic reconstruction. Each workflow receives the same initial human motion, object mesh, object trajectory, and source video. We compare the resulting full-sequence reconstructions before physical refinement, rather than the annotations in isolation. For each output, annotation time sums the hands-on editing operations across the sequence. We average the two annotators' results within each sequence and then average over the 20 sequences.

D.2 Comparison Workflows

Blender-like object-pose adjustment. Fig. 14(a) shows the synchronized RGB and 3D views used for direct object-pose editing. At a frame where contact changes, the annotator translates and rotates the object to match the video and its interaction with the reconstructed human. The interface computes nearby human and object surface regions from this pose using a distance threshold. The annotator can inspect the inferred region, revise the pose, and recompute it before saving the event. The resulting region replaces the contact region in our agent-based HOI optimization; it remains unchanged until the next annotated contact change. The manual pose edit is therefore used to specify contact, not as a separate framewise reconstruction result.

Painted contact regions. Fig. 14(b) shows the surface-brush interface. At contact-change frames, the annotator paints the contacting areas on the human and object meshes while inspecting the source video; add and erase controls allow the regions to be revised. The selected regions are then carried through the interval until the next contact change. As in the Blender-like workflow, the regions remain unchanged within that interval and replace the contact region in the same agent-based HOI optimization. This workflow specifies contact directly on the two surfaces, without first using an adjusted object pose to derive it.

Sparse point pairs. Our interface instead records corresponding human and object contact points, together with object-to-image observations. The point tracks are reused while the corresponding contact remains stable, and annotators update them when it changes. Point pairs specify the human and object locations that should meet, providing stronger geometric constraints than painted regions, which leave the surface correspondence ambiguous. Painted regions can also become outdated between annotations when contact shifts across the surfaces.



(a) Blender-like object-pose adjustment (b) Painted contact-region annotation

Figure 14: Interfaces for the two manual-repair baselines. Both show the source video beside an interactive 3D reconstruction. The Blender-like tool adjusts object pose and infers contact regions from surface distance; the surface-brush tool marks human and object contact regions directly. In both cases, the annotated region is used by the same downstream agent-based HOI optimization.

D.3 Human Quality Review

We report penetration-depth P95 separately, using the geometric evaluation protocol in Sec. B.3. Five independent reviewers, who were neither method developers nor the two annotators, assessed rendered reconstruction videos for all three workflows, using the source videos as references. Method names were hidden and presentation order was randomized. Each reviewer followed the

instructions below to produce a weighted score out of 100. Scores were averaged across reviewers, the two annotators' outputs, and the 20 sequences. Failure is judged by evaluators uninvolved in development, with method names hidden and result order randomized. An output fails if it shows obvious object tracking failure, such as severe interpenetration or a physically implausible pose. The instructions provided to reviewers were:

> *Compare the three results within each case, then score each method from 0 to 100 using:*
>
> - *Contact plausibility (35%): whether the intended body part contacts the appropriate object region, without visible gaps, missed contacts, or contact with the wrong surface.*
> - *Physical plausibility (30%): visible interpenetration, floating, loss of support, or implausible human–object spatial relationships.*
> - *Temporal consistency (10%): whether contact and object placement evolve coherently throughout the rendered video, without visible jitter or abrupt changes.*
> - *Source-action agreement (25%): whether the reconstruction preserves the basic action shown in the source video.*
>
> *Score each aspect from 0 to 100 using these anchors: 0 = persistent severe failure; 20 = repeated clear errors; 40 = substantial errors; 60 = mixed quality; 80 = minor errors; 100 = no visible error. Intermediate scores are permitted. Compute the final score using the weights above. Repeated obvious contact or physical errors must receive low scores on the corresponding aspects. Reflect clear quality differences, but do not force a particular method to win or impose a minimum score gap.*
>
> *Do not score pixel similarity, texture, lighting, background, or camera-view differences. Judge only visible evidence in the videos and do not infer unobserved forces. If an aspect cannot be judged reliably, exclude it and renormalize the remaining weights.*
>
> *Record the four aspect scores and the final weighted score for each result.*

E More Results

Fig. 15 provides additional reconstruction comparisons with CARI4D on diverse interactions. In these retained examples, both methods use the same input videos and our reconstructed object meshes, while CARI4D retains its original human estimation and interaction optimization. This shared-mesh qualitative setting should be distinguished from the published InterCAP comparison. The examples illustrate relative human–object placement and contact over time.

Fig. 16 shows additional examples of physical refinement across different human–object interactions.

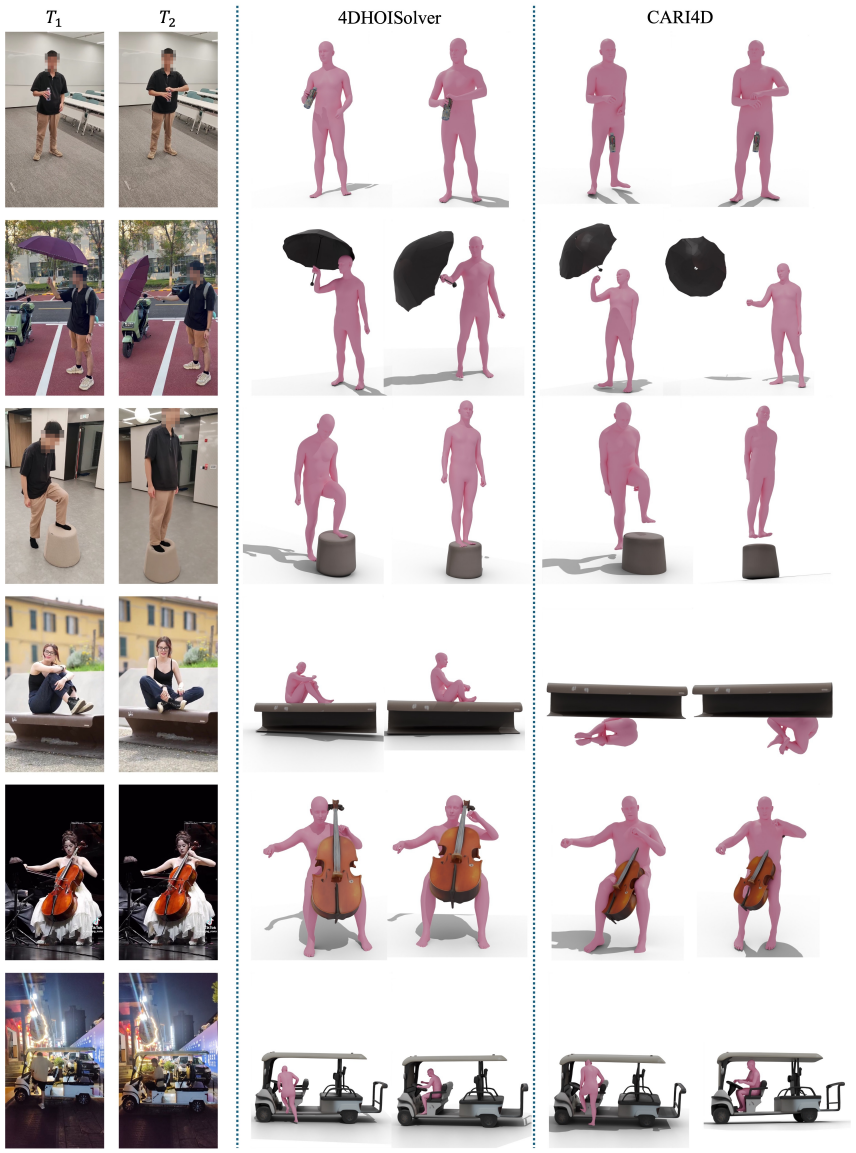


Figure 15: Additional qualitative comparisons with CARI4D. Each row shows two input frames (T_1, T_2) and the corresponding reconstructions from 4DHOISolver and CARI4D. The examples illustrate human–object placement and contact across diverse interactions, using shared object geometry.

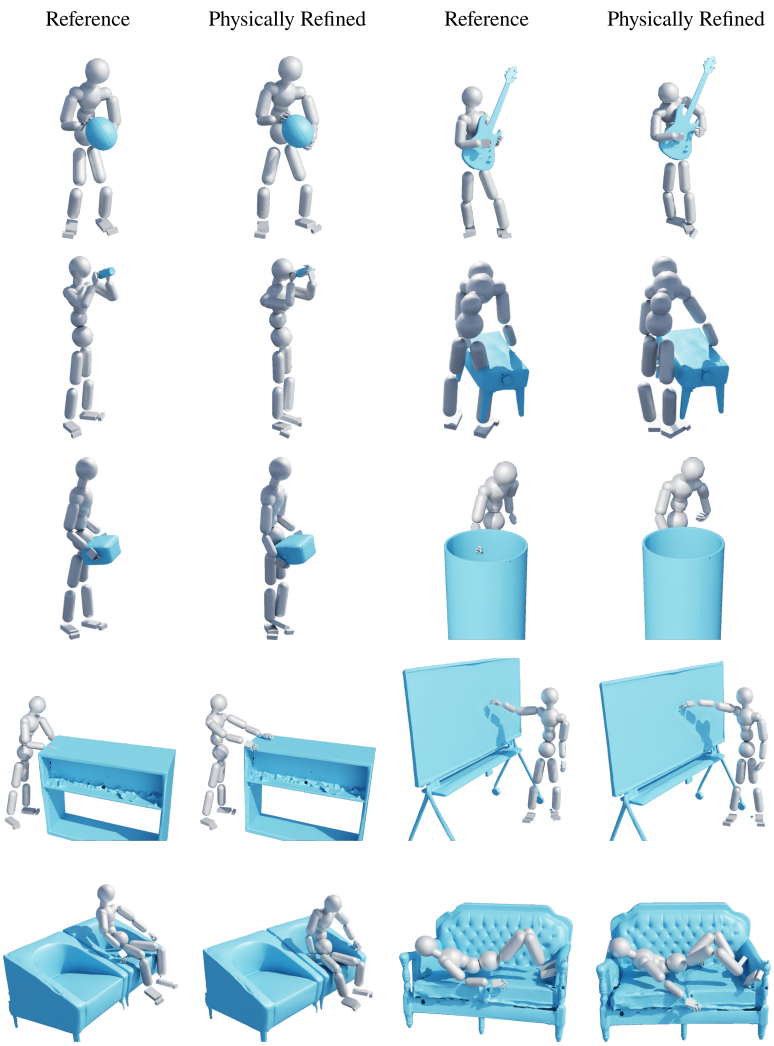


Figure 16: Additional qualitative results of physical refinement. Each adjacent pair compares the input reference with its physically refined motion across diverse human–object interactions.