Real2Sim in HOI: Toward Physically Plausible HOI Reconstruction from Monocular Videos

Yubo Zhao[1,2] **Yujin Chai**[2] **Yunao Dong**[1] **Chengfeng Zhao**[1]
Zijiao Zeng[2] **Yuan Liu**[1] **Chi-Keung Tang**[1]
[1]The Hong Kong University of Science and Technology [2]Tencent Games Motus

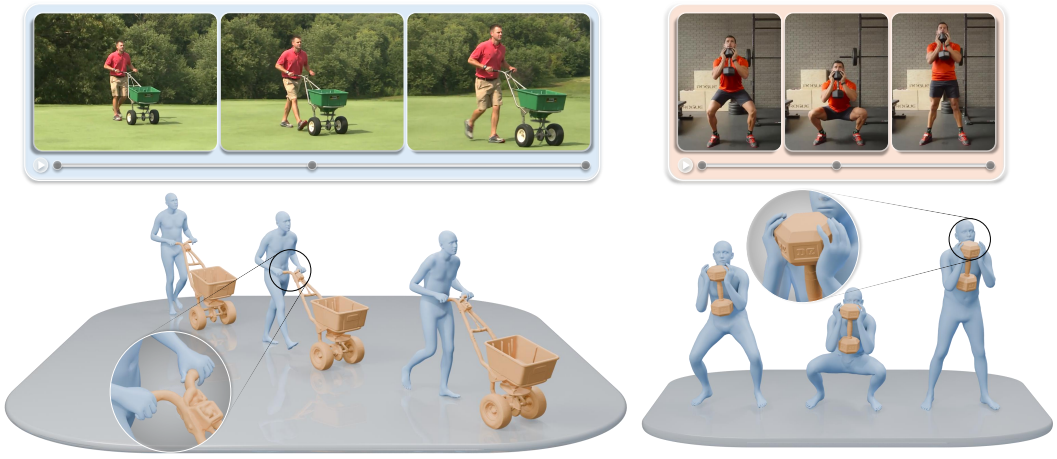

Figure 1: **Reconstructing interaction, not just trajectories.** From monocular RGB videos, *HA-HOI* reconstructs simulation-ready 4D human-object interaction.

Abstract

Recovering 4D human-object interaction (HOI) from monocular video is a key step toward scalable 3D content creation, embodied AI, and simulation-based learning. Recent methods can reconstruct temporally coherent human and object trajectories, but these trajectories often remain visual artifacts while failing to preserve stable contact, functional manipulation, or physical plausibility when used as reference motions for humanoid-object simulation. This reveals a fundamental interaction gap: HOI reconstruction should not stop at tracking a human and an object, but should recover the relation that makes their motion a coherent interaction. We introduce *HA-HOI*, a framework for reconstructing physically plausible 4D HOI animation from in-the-wild monocular videos. Instead of treating the human and object as independent entities in an ambiguous monocular 3D space, we propose a *human-first, object-follow* formulation. The human motion is recovered as the interaction anchor, and the object is reconstructed, aligned, and refined relative to the human action. The resulting kinematic trajectory is then projected into a physics-based humanoid-object simulation, where it acts as a teacher trajectory for stable physical rollout. Across benchmark and in-the-wild videos, *HA-HOI* improves human-object alignment, contact consistency, temporal stability, and simulation readiness over prior monocular HOI reconstruction methods. By moving beyond visually plausible trajectory recovery toward physically grounded interaction animation, our work takes a step toward turning general monocular HOI videos into scalable demonstrations for humanoid-object behavior. [`Project page`]

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

4D human-object interaction (HOI) reconstruction aims to recover temporally coherent and spatially consistent motions of humans and objects from visual observations. Converting videos into animatable human-object motion, it provides a key foundation for 3D content creation, AR/VR, embodied AI, and simulation-based learning. In particular, reinforcement learning has greatly benefited from kinematic human motion data as reference demonstrations for acquiring human-like behaviors (Peng et al., 2018, 2021; Xu et al., 2025; Wang et al., 2026). However, obtaining reliable 4D HOI animation remains difficult. Traditional pipelines typically rely on human and object motion capture, multi-view RGB-D systems, or extensive manual post-processing. While effective in controlled environments, such setups are expensive, labor-intensive, and difficult to scale.

Monocular RGB videos, in contrast, are abundant, easy to capture, and represent the most natural form in which human-object interactions are recorded in the wild. They offer a promising path toward scalable HOI reconstruction from everyday videos. Yet monocular reconstruction is fundamentally under-constrained.

Recent works such as VisTracker (Xie et al., 2023), CARI4D (Xie et al., 2026), and THO (Zhang et al., 2026) have made important progress in monocular human-object tracking and reconstruction. These methods can recover human and object trajectories in 3D space, and some further introduce contact-aware modules or learned interaction constraints (Wang et al., 2025c). Nevertheless, their outputs often remain visually plausible 4D tracks rather than interaction motions that are physically stable, transferable, or directly usable as humanoid-object animation. We identify three limitations that are especially important for this stricter goal. First, most existing methods are designed around image- or camera-space reconstruction, and are not explicitly organized around a stable, gravity-aware interaction frame, which becomes crucial for moving-camera in-the-wild videos. Second, although contact cues are introduced, they are often coarse, latent, or unreliable, and existing pipelines do not explicitly recover fine hand-object articulation as a first-class part of the interaction. Third, these methods are primarily evaluated as reconstruction systems rather than as teacher trajectories for physical humanoid-object rollout. As a result, even when the recovered human and object trajectories appear reasonable, they may fail to preserve stable contact, functional manipulation, or dynamic plausibility when instantiated in a physical humanoid-object system.

We argue that this formulation misses the central structure of HOI. Human-object interaction is not merely the coexistence of a human trajectory and an object trajectory in 3D space. It is defined by the evolving functional relation between them. We identify this as an *interaction gap* in current monocular HOI reconstruction pipelines. Although they may recover human and object trajectories, they often fail to recover the physical and semantic relation that binds those trajectories into a coherent interaction. Without such articulated interaction, reconstructed HOI animations remain limited for downstream applications that require contact, affordance, and action-level consistency.

To address this gap, we present *HA-HOI* (short for **H**uman-**A**nchored **HOI**), a framework for reconstructing physically plausible 4D HOI animation from in-the-wild monocular videos. Our goal is not merely to recover a visually aligned human-object trajectory, but to produce an interaction sequence that can guide a simulated humanoid toward reasonable object-interaction behavior. Our key insight is that, in most HOI videos, the human is the active agent and the object motion is conditioned on human action. Therefore, the human body should not be treated as just another reconstructed entity in the scene; it should serve as the organizing coordinate system of the interaction. We formulate monocular HOI reconstruction as a *human-first, object-follow* process.

Specifically, instead of reconstructing the human and object as two independent trajectories in an ambiguous monocular 3D space, we first recover a coherent human motion and use it as the anchor of the interaction. The object is then initialized, constrained, and refined relative to the human motion, where contact, affordance, and interaction can be expressed more naturally. We further incorporate a VLM-based interaction model to propose likely contact surfaces, which allows recovering more meaningful body-object relations while keeping the reconstruction grounded by visual alignment, temporal consistency, and geometric constraints. Finally, we use the reconstructed sequence as a teacher trajectory for a humanoid controller, requiring the output to not only look plausible in 3D, but also support stable contact and reasonable object interaction under physical rollout.

Empirically, *HA-HOI* reconstructs more coherent and simulation-ready HOI animations from in-the-wild monocular videos. Compared with prior monocular HOI reconstruction methods, our framework

improves human-object alignment, contact consistency, temporal stability, and physical rollout quality. Our results suggest that scalable 4D HOI reconstruction should move beyond independent entity tracking and instead treat interaction itself as the organizing principle.

In summary, our contributions are:

- We introduce a scalable and flexible framework for transforming in-the-wild monocular HOI videos into simulation-ready humanoid-object interaction trajectories, moving beyond visually aligned 4D tracking toward physically grounded interaction animation, where:
- A human-first, object-follow reconstruction pipeline *HA-HOI* is proposed that uses world-grounded human motion as the interaction anchor, reconstructs object motion relative to the human action, and refines fine-grained body-object contact through VLM-guided contact proposals and explicit hand/body optimization.
- We demonstrate that *HA-HOI* achieves accurate 4D HOI reconstruction while significantly improving physical plausibility, contact consistency, and temporal stability for humanoid-object simulation.

2 Related Work

2.1 3D Foundation Models

Recent 3D foundation models have greatly advanced monocular reconstruction for humans, objects, and scenes, offering strong priors for body pose, hand geometry, object shape, depth, camera parameters, point maps, tracks, and object pose from in-the-wild images and videos (Goel et al., 2023; Pavlakos et al.; Shen et al., 2024; Shin et al., 2024; Xia et al., 2025; Zhao et al., 2025; Xiang et al., 2025; Xu et al., 2024; Wu et al., 2024; Yang et al., 2026; Wang et al., 2025a,b, 2024; Wen et al., 2024; Labbé et al., 2022; Cao et al., 2025; Li et al., 2026; Lu et al., 2026). Yet 4D dynamic scene recovery remains highly challenging. Monocular reconstruction is inherently ambiguous, temporal tracking amplifies geometric inconsistency, and most foundation models are still designed for humans, objects, or scenes in isolation. Consequently, directly composing these priors often produces inconsistent geometry, unstable motion, and weak human-object contact, making temporally and physically consistent 4D HOI reconstruction far from solved.

2.2 3D HOI Reconstruction

Static 3D human-object interaction reconstruction has been studied in both hand-object and full-body settings. Existing methods span learning-based joint prediction (Hasson et al., 2019), hybrid contact-aware estimation (Yang et al., 2021), and optimization-based reconstruction in the wild (Cao et al., 2021) for hand-object interaction, as well as spatial fitting (Zhang et al., 2020), joint neural reconstruction (Xie et al., 2022), contact-driven fitting (Cseke et al., 2025), foundation-model-guided interaction reasoning (Dwivedi et al., 2025), large-model-guided reconstruction (Liu et al., 2025), and diffusion-based reconstruction (Li et al., 2025) for more general human-object interaction. Their development is also supported by increasingly diverse hand-object benchmarks (Shivakumar et al., 2019; Fan et al., 2023). However, many existing methods still rely on supervised training or fitted priors from relatively limited HOI data collected in controlled settings, which restricts generalization to unconstrained in-the-wild scenes. Moreover, although some methods introduce explicit contact losses or contact priors to improve physical plausibility (Hasson et al., 2019; Yang et al., 2021; Dwivedi et al., 2025; Li et al., 2025), they remain largely image-based or frame-wise and do not explicitly enforce temporal consistency, limiting their applicability to stable 4D HOI reconstruction.

2.3 4D HOI Tracking

For 4D HOI, recent progress has been enabled by datasets and benchmarks for dynamic human-object capture (Bhatnagar et al., 2022; Huang et al., 2024; Zhang et al., 2024). Building on these resources, existing methods recover human and object motion from monocular videos through template-based tracking, optimization-based refinement, and transformer-based temporal modeling (Xie et al., 2023; Zhao et al., 2024; Xie et al., 2025; Zhang et al., 2026; Xie et al., 2026). Despite these advances, explicit interaction modeling remains limited: most methods still rely on coarse geometric cues, and even contact-aware approaches mainly use contact to reduce mesh distance and penetration

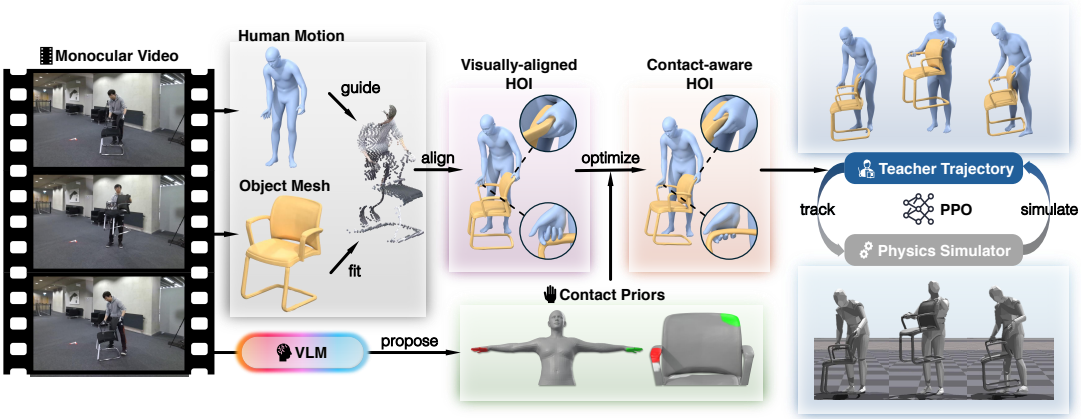

Figure 2: **Overview of *HA-HOI*.** From a monocular HOI video, our pipeline recovers human motion and object geometry, aligns them into a visually consistent 4D HOI sequence, and uses VLM-proposed contact priors to refine body-object interaction. The resulting physically plausible HOI trajectory serves as a teacher motion for humanoid-object simulation, enabling stable physical rollout from in-the-wild video.

rather than recover fine-grained contact geometry. Moreover, current 4D HOI methods generally do not incorporate physical simulation, making it difficult to obtain physically grounded interaction dynamics. Although physics-based tracking has been explored for human motion (Tevet et al., 2024), it has not been extended to human-object interaction. Motivated by these limitations, we propose a training-free pipeline for more temporally consistent and physically plausible monocular 4D HOI reconstruction.

3 Methodology

We present *HA-HOI*, a framework that reconstructs physically plausible 4D human-object interaction from a monocular RGB video. Given an input video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, our goal is to recover an interaction trajectory $\tau = \{H_t, O_t\}_{t=1}^{T}$ that is visually aligned with the video, temporally coherent, contact-consistent, and suitable as a teacher trajectory for physical humanoid-object rollout.

We represent the human state at frame t as $H_t = (\boldsymbol{\beta}, \boldsymbol{\theta}_t, \boldsymbol{\psi}_t, \mathbf{R}_t^h, \mathbf{t}_t^h)$, where $\boldsymbol{\beta}$ denotes body shape, $\boldsymbol{\theta}_t$ body pose, $\boldsymbol{\psi}_t$ hand pose, and $(\mathbf{R}_t^h, \mathbf{t}_t^h)$ global body orientation and translation. The object is represented by a metric mesh $\mathcal{M}_o$ and a 6-DoF pose sequence $O_t = (\mathbf{R}_t^o, \mathbf{t}_t^o) \in SE(3)$.

The key design of *HA-HOI* is to reconstruct interaction in a human-centric manner. Instead of independently reconstructing a human track and an object track in an ambiguous monocular 3D space, we first establish the human motion as the interaction anchor. The object is then recovered relative to this anchor, refined using semantic contact proposals, and finally projected into a physical humanoid-object simulation. This yields a pipeline with four stages: human-centric anchoring, object-follow reconstruction, semantic contact refinement, and physics-based interaction projection.

3.1 Human-Centric Interaction Anchor

Monocular depth and point clouds provide useful local geometric cues, but they are often not stable enough to define the full 4D scene, especially under occlusion and camera motion. We therefore begin by constructing a temporally coherent human motion and use it as the coordinate system of the interaction.

To instantiate this human anchor, we recover world-space body motion represented by SMPL-X (Pavlakos et al., 2019; Shen et al., 2024). Since hand articulation is critical for human-object interaction, we further estimate MANO hand poses and fuse them into the SMPL-X hand parameters (Romero et al., 2017; Pavlakos et al.). This gives an articulated human sequence $\{H_t^0\}_{t=1}^{T}$ that provides both global body motion and fine-grained hand motion.

We then calibrate monocular depth to this human anchor. Let D_t be the monocular depth prediction (Wang et al., 2025b) at frame t. Rather than treating the depth map as an independent global reconstruction, we estimate a scale factor that aligns the unprojected depth points in the human region

with the recovered human mesh. The calibrated depth serves as local geometric evidence for object placement while preserving the temporally coherent scale of the human motion. In this way, the human body defines the metric interaction frame, and depth is used as supporting evidence rather than as the sole scene representation.

3.2 Object-Follow Reconstruction

Given the human-centric frame, we next recover the manipulated object as an object-follow trajectory. The purpose of this stage is not to solve general object reconstruction in isolation, but to obtain an object geometry and pose sequence that remain visually grounded in the video and compatible with the human motion.

We first construct a canonical object mesh $\mathcal{M}_o^c$ from selected frames in which the object is sufficiently visible. When the object is partially occluded by the human, we use image inpainting to obtain a cleaner object reference before applying a single-image 3D object reconstruction model (Google, 2026; Zhao et al., 2025). Since monocular object reconstruction does not reliably determine metric scale, we estimate an object scale $s_o^\star$ using reference frames with reliable masks and calibrated depth. For each candidate scale s, we align the object to the observed image, render its silhouette and depth, and compare them with the observed object mask and calibrated depth. We choose the scale with the best agreement across reference frames:

$$s_o^\star = \arg\max_s \text{median}_{t\in\mathcal{T}_{\text{ref}}} \left[\text{IoU}(\hat{S}_t(s), S_t) \exp\left(-\frac{\Delta z_t(s)^2}{2\sigma_d^2} \right) \right]. \tag{1}$$

Here the first term measures silhouette agreement and the second term penalizes depth mismatch after calibration to the human anchor. The median over reference frames makes the estimate robust to occasional occlusion or tracking failure. The metric object mesh is then $\mathcal{M}_o = s_o^\star \mathcal{M}_o^c$.

With the scaled object mesh, we track the object pose across the video using FoundationPose (Wen et al., 2024). Since object tracking can fail under occlusion or motion blur, we use mask-depth consistency as a reliability signal. Poses with poor agreement are treated as unreliable and are recovered through interpolation or re-initialization from nearby reliable frames. Finally, the tracked camera-space object poses are transformed into the human-centric world frame.

This stage produces a coarse HOI trajectory

$$\tau^0 = \{H_t^0, O_t^0\}_{t=1}^T.$$

Although this trajectory is visually grounded and temporally coherent, it may still fail to express the actual interaction: the object may float near the hand, penetrate the body, or touch an incorrect region while still giving the same visual at capture viewpoint. We therefore refine it using semantic contact reasoning.

3.3 Semantic Contact Refinement

The coarse trajectory provides human and object motion, but not necessarily a coherent body-object relation. To convert this trajectory into an interaction trajectory, we introduce semantic contact refinement. We use InteractVLM (Dwivedi et al., 2025) as a contact proposal module. Given representative frames of an interaction segment, it predicts likely contact regions on both the human and the object. These predictions are mapped to vertex sets $\mathcal{C}_h \subset \mathcal{M}_h, \mathcal{C}_o \subset \mathcal{M}_o$, where $\mathcal{M}_h$ is the SMPL-X human mesh and $\mathcal{M}_o$ is the reconstructed object mesh.

The final interaction is determined by optimizing agreement among contact cues, visual alignment, temporal smoothness, and physical regularity. Starting from τ^0, we optimize only the variables that directly affect interaction:

$$\mathcal{X} = \{\boldsymbol{\theta}_t^{\text{sel}}, \boldsymbol{\psi}_t, \mathbf{R}_t^o, \mathbf{t}_t^o\}_{t=1}^T,$$

where $\boldsymbol{\theta}_t^{\text{sel}}$ denotes selected body joints, $\boldsymbol{\psi}_t$ the hand pose, and $(\mathbf{R}_t^o, \mathbf{t}_t^o)$ the object pose.

We minimize

$$\min_{\mathcal{X}} \sum_{t=1}^{T} \left(\lambda_c \mathcal{L}_c + \lambda_p \mathcal{L}_p + \lambda_s \mathcal{L}_s + \lambda_a \mathcal{L}_a + \lambda_r \mathcal{L}_r \right), \tag{2}$$

where $\mathcal{L}_c$ encourages contact consistency, $\mathcal{L}_p$ penalizes penetration, $\mathcal{L}_s$ encourages temporal smoothness, $\mathcal{L}_a$ anchors the refinement to the coarse reconstruction, and $\mathcal{L}_r$ regularizes body and hand poses.

The core term is the contact loss. During early iterations, we use a bidirectional nearest-neighbor distance between predicted human and object contact regions to obtain stable coarse contact. Once the interaction is approximately aligned, we switch to an object signed-distance-field contact objective:

$$\mathcal{L}_c = \sum_{s\in\{l,r\}} \frac{1}{|\mathcal{C}_h^s|} \sum_{i\in\mathcal{C}_h^s} \Phi_o\left((\mathbf{R}_t^o)^\top(\mathbf{x}_{t,i}^h - \mathbf{t}_t^o)\right)^2, \tag{3}$$

where $\mathbf{x}_{t,i}^h$ is a human contact vertex and Φ_o is the signed distance field of the object in its canonical coordinate frame. This objective pulls predicted human contact vertices onto the object surface while allowing the object and local hand/body pose to adapt.

Our progressive optimization proceeds as follows: First, we refine the SMPL translation under depth and contact constraints to improve human-object alignment. Next, we optimize the object's 6-DoF pose while enforcing contact consistency, penetration constraints, and temporal regularization. We then jointly refine the human body pose and object trajectory to resolve their remaining geometric inconsistencies. Finally, we optimize articulated hand poses using contact, penetration, temporal regularization, and hand-pose priors.

The result is a contact-refined kinematic trajectory

$$\tau^\star = \{H_t^\star, O_t^\star\}_{t=1}^T,$$

which is visually aligned and geometrically interaction-aware. However, it is still a kinematic reconstruction. A trajectory can satisfy geometric contact losses while remaining unstable under gravity or rigid-body dynamics. We therefore use physical simulation as a final interaction projection step.

3.4 Physics-Based Interaction Projection

We use physics-based tracking as the final projection step by instantiating a corresponding humanoid-object system in a physics simulator. We convert the optimized human sequence $H_t^\star$ into a humanoid reference motion with generalized coordinates $\bar{\mathbf{q}}_t$, and use the optimized object sequence $O_t^\star$ as the reference object trajectory $\bar{O}_t$. The object is modeled as a dynamic rigid body, and the humanoid is controlled through joint-level actuation.

Rather than directly replaying the kinematic reconstruction, we train a sequence-specific residual controller. At simulation step t, the policy observes the current simulated state and the reference state in a root-relative coordinate frame. It outputs a residual action $\mathbf{a}_t$, which modifies the reference joint target:

$$\mathbf{q}_t^{\mathrm{tar}} = \bar{\mathbf{q}}_t + \alpha\mathbf{a}_t,$$

where α controls the maximum deviation from the reconstructed reference. The humanoid is driven toward $\mathbf{q}_t^{\mathrm{tar}}$ using PD control, while the object evolves under contact and gravity.

This formulation treats the reconstructed trajectory as a strong teacher while allowing the simulator to correct small errors that would otherwise cause slipping, jitter, or unstable contact. The tracking reward balances motion fidelity and interaction stability:

$$r_t = w_h r_t^h + w_o r_t^o + w_{\mathrm{hand}} r_t^{\mathrm{hand}} + w_{\mathrm{contact}} r_t^{\mathrm{contact}} + w_{\mathrm{vel}} r_t^{\mathrm{vel}}. \tag{4}$$

The human tracking term encourages the simulated humanoid to follow the reconstructed body motion. The object term rewards agreement with the reference object trajectory. The hand term emphasizes grasp-related articulation. The contact term encourages predicted body-object contacts to remain active during rollout, and the velocity term discourages implausible accelerations.

After training, we roll out the controller once and convert the simulated states back into SMPL-X parameters and object poses. The final output is

$$\tau^{\mathrm{phys}} = \{H_t^{\mathrm{phys}}, O_t^{\mathrm{phys}}\}_{t=1}^T,$$

a physically plausible humanoid-object animation that remains close to the input video while supporting stable interaction under simulation.

Table 1: Quantitative comparison on BEHAVE. Reconstruction quality is evaluated using Chamfer distance (CD, cm) and acceleration error (Acc, cm/s^2). Interaction quality is assessed through penetration (Pen., cm), contact recall (C-Recall, %), rollout coverage (RC, %), and contact retention (CR, %). The last two metrics are evaluated under physics-based rollout. Our method requires no ground-truth object template at test time. *CARI4D is evaluated on 62 selected sequences.

Method	Moving Camera	Template-free	Reconstruction Quality					Interaction & Physical Plausibility			
			CD-h↓	CD-o↓	CD-c↓	Acc-h↓	Acc-o↓	Pen.↓	C-Recall↑	RC↑	CR↑
CHORE	✗	✗	25.82	28.94	26.52	2.75	4.59	–	–	–	–
InterTrack	✗	✓	25.71	47.66	30.20	5.23	5.64	–	–	–	–
HOI-TG	✗	✗	21.19	24.70	22.67	2.44	3.76	–	–	–	–
VisTracker	✗	✗	11.24	12.58	11.63	0.54	0.77	0.079	27.5	9.2	37.9
CARI4D*	✗	✓	7.51	**11.93**	**9.06**	1.07	**0.41**	0.084	34.5	10.8	23.1
THO	✗	✗	8.49	13.36	10.06	0.79	1.02	–	–	–	–
Ours	✓	✓	**7.11**	15.68	10.74	**0.52**	0.59	**0.013**	**37.9**	**31.4**	**85.0**

4 Experiments

4.1 Experiments Setup

Datasets. We evaluate our method on the BEHAVE dataset (Bhatnagar et al., 2022), a full-body human-object interaction benchmark captured with multi-view RGB-D sensors and released with pseudo-ground-truth SMPL body and object registrations. Following the standard monocular HOI reconstruction protocol used by prior work (Xie et al., 2023, 2026; Zhang et al., 2026), we use the extended BEHAVE annotations, where human and object registrations are provided at 30 FPS. We adopt the official split with 217 training sequences and 82 test sequences, and report results on the test set. Although BEHAVE is captured with multi-view RGB-D sensors, our method takes monocular RGB video as input.

Metrics and Evaluation Protocol. We assess reconstruction quality using Chamfer distance (CD) between predicted and ground-truth surfaces for the human (CD-h), object (CD-o), and their combination (CD-c), together with acceleration error (Acc-h, Acc-o) to measure temporal consistency. To evaluate interaction quality, we report human-object penetration (Pen.), measured as the mean penetration depth of human mesh vertices into the object and contact recall (C-Recall), which measures the fraction of ground-truth contacts recovered within a 1.5 cm threshold. We further assess physical executability through rollout coverage (RC), defined as the longest valid continuous rollout relative to the reference sequence length, and contact retention (CR), the fraction of intended-contact frames that maintain contact during simulation.

All methods are evaluated from monocular RGB input. Our pipeline requires no BEHAVE-specific training, ground-truth object templates, pose initialization, contact annotations, or sequence-level manual adjustment. Reconstruction and contact metrics are computed before physics simulation, while RC and CR are measured under a controlled physics-projection protocol with identical simulator settings and policy-training configurations. Following prior 4D HOI evaluation protocols (Xie et al., 2023, 2026), we align the predicted human mesh to ground truth in the first frame and apply the resulting transformation to the entire sequence, avoiding per-frame alignment that could conceal temporal drift. We convert SMPL-X reconstructions to SMPL-H for compatibility with ground-truth annotations and prior baselines.

4.2 Quantitative Comparison on BEHAVE

Table 1 compares *HA-HOI* with recent monocular HOI reconstruction and tracking methods, including CHORE (Xie et al., 2022), InterTrack (Xie et al., 2025), HOI-TG (Wang et al., 2025c), VisTracker (Xie et al., 2023), CARI4D (Xie et al., 2026), and THO (Zhang et al., 2026).

Reconstruction Quality. Different from many prior methods that rely on a ground-truth object template at test time or learn BEHAVE-specific interaction priors, our method runs the full pipeline end-to-end from monocular RGB video. It does not assume a predefined object mesh, object pose initialization, contact annotation, or sequence-level manual adjustment. Despite this general setting, our method achieves competitive geometric accuracy and temporal consistency, obtaining the lowest human Chamfer distance (7.11 cm) and acceleration error (0.52 cm/s^2).

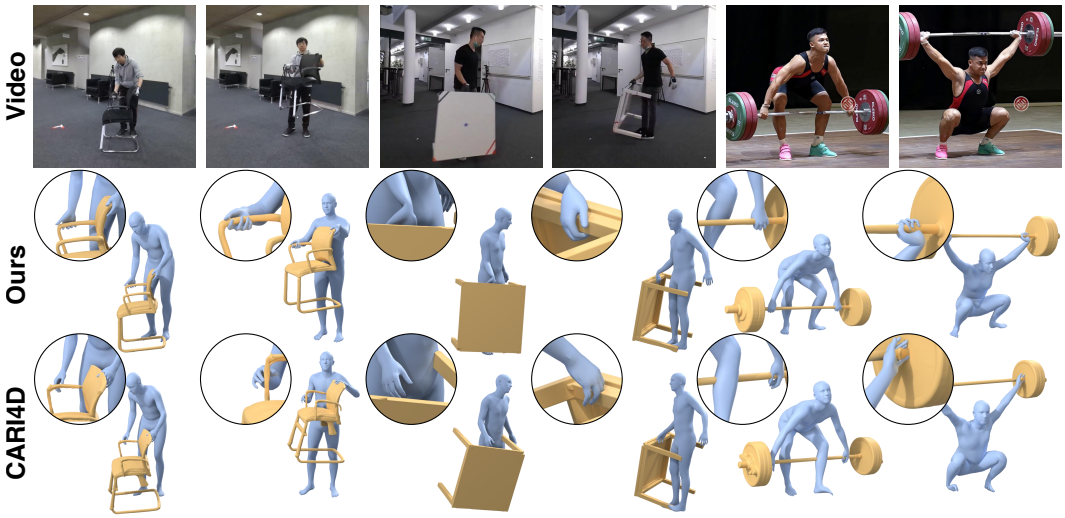

Figure 3: Qualitative comparison with CARI4D on BEHAVE test sequences, BEHAVE training sequences, and in-the-wild videos. CARI4D often fails to maintain functional hand-object relations, leading to floating, missing, or ambiguous contacts. In contrast, *HA-HOI* reconstructs contact-aware HOI animations with articulated hands, stable object support, and more physically plausible interaction geometry.

Interaction Quality and Physical Executability. More importantly, the advantages of *HA-HOI* become substantially more pronounced when evaluating the recovered interaction rather than individual reconstructed surfaces. Our method achieves the lowest human-object penetration while attaining the highest contact recall, demonstrating improved geometric compatibility without sacrificing contact fidelity. These advantages become particularly evident under physical rollout. With identical simulation and policy-training settings, *HA-HOI* achieves 31.4% rollout coverage and 85.0% contact retention, substantially outperforming baselines.

Notably, the modest improvement in contact recall contrasts sharply with the substantial gains in rollout coverage and contact retention. This suggests that the physical executability of an interaction depends not only on recovering correct contact locations, but also on establishing a coherent relationship between human motion and object trajectories. Rather than treating contact as isolated surface correspondences, *HA-HOI* jointly refines human and object poses under semantic guidance, encouraging a mutually compatible interaction throughout the sequence. The results demonstrate that *HA-HOI* recovers not merely geometrically plausible contacts, but more sustainable human-object interactions that provide stronger teacher trajectories for physics-based animation.

Generalization to InterCap. We further evaluate *HA-HOI* on InterCap (Huang et al., 2024) to examine its generalization beyond BEHAVE. As shown in Table 2, our method achieves the best human and combined reconstruction accuracy, together with the lowest human and object acceleration errors. These consistent improvements across both benchmarks support the effectiveness of our human-anchored formulation beyond a single dataset.

Table 2: Quantitative comparison on InterCap. Chamfer distance (CD, cm) and acceleration error (Acc, cm/s^2).

Method	CD-h↓	CD-o↓	CD-c↓	Acc-h↓	Acc-o↓
VisTracker	16.12	27.41	20.17	0.98	1.25
CARI4D	11.06	**15.69**	12.88	1.25	0.82
Ours	**8.69**	17.12	**12.76**	**0.75**	**0.49**

4.3 Qualitative Results

Figure 3 provides qualitative comparisons on three representative videos: a BEHAVE test sequence, a BEHAVE training sequence, and an in-the-wild interaction video. Our results are shown after physics simulation, which makes the comparison more demanding than visual reconstruction alone.

HA-HOI produces stable, temporally coherent, and visually aligned HOI animations across all three settings. In the chair and table examples, our method maintains plausible hand-object contact and object placement throughout the interaction. In the weightlifting example, our reconstruction

Table 3: Ablation study on BEHAVE evaluating the effects of ground-truth object geometry, camera parameters, contact optimization, and contact guidance.

Method	CD-h↓	CD-o↓	Acc-h↓	Acc-o↓	Pen.↓	C-Recall↑	RC↑	CR↑
Ours	7.11	15.68	0.52	0.59	0.013	37.9	31.4	85.0
w/ GT mesh	7.13	13.96	0.52	0.58	0.007	37.4	31.6	86.4
w/ GT cam.	7.08	15.19	0.52	0.57	0.014	38.2	31.5	84.9
w/o contact opt.	7.10	15.49	0.51	0.49	0.097	31.2	7.2	24.2
w/o guidance	7.08	15.44	0.51	0.52	0.011	33.6	10.2	38.5

preserves the grasp-like relation between both hands and the barbell, allowing the object to remain physically tied to the human action after simulation.

Compared with CARI4D, the advantage of *HA-HOI* is most visible around contact regions. CARI4D often reconstructs a visually-aligned global human-object layout from camera view, but the body parts do not reliably explain the interaction: the object may be close to the body while the fingers float, penetrate, or fail to establish stable contact. This is a critical failure mode for previous 4D HOI reconstruction, because the interaction is not defined only by where the human and object are, but by how the human manipulates the object.

These examples further illustrate how *HA-HOI* preserves the functional structure of human-object interaction across diverse scenarios. In particular, the articulated hand configurations and coordinated object motion reveal interaction details that are difficult to appreciate from aggregate quantitative results alone.

4.4 Ablation Study

Table 3 investigates the contributions of object geometry, camera estimation, contact optimization, and semantic contact guidance. Replacing the estimated object mesh or camera parameters with ground truth yields only modest improvements, and the resulting rollout coverage and contact retention remain largely unchanged, suggesting that neither component is the dominant bottleneck in interaction reconstruction. *HA-HOI* remains stable under the fully automatic setting, where neither the ground-truth object mesh nor the ground-truth camera is provided.

The contact ablations further reveal the importance of interaction-aware refinement. Removing the entire contact optimization stage increases penetration and reduces rollout coverage and contact retention dramatically. To further examine the role of semantic contact guidance within this optimization, we remove the VLM-proposed contact regions while retaining the remaining refinement objectives. Interestingly, this variant maintains comparable reconstruction accuracy and even slightly lower penetration (0.011 cm), yet rollout coverage and contact retention decline substantially to 10.2% and 38.5%. This suggests that geometric refinement alone cannot fully recover the functional relationships needed to sustain human-object interaction. Semantic contact guidance provides an additional constraint on how the human and object should interact, allowing the joint optimization to recover more coherent body articulation and object motion.

Together, these results demonstrate that the effectiveness of contact optimization arises not merely from reducing geometric inconsistencies, but from incorporating interaction-level guidance into the coupled refinement of human and object motion. This distinction is central to *HA-HOI*: recovering a physically sustainable interaction requires more than geometrically compatible trajectories.

5 Conclusion

We presented *HA-HOI*, a framework for reconstructing physically plausible 4D human-object interactions from monocular videos. Our central insight is that a coherent interaction emerges not merely from accurate human and object trajectories, but from the functional relation that couples their motions. By explicitly recovering this relation, *HA-HOI* achieves strong reconstruction accuracy while substantially improving contact consistency and physical executability. More broadly, our work points to a new direction for monocular 4D HOI reconstruction: using ordinary videos not only as visual observations, but as scalable sources of interaction demonstrations. We believe this shift from trajectory reconstruction to physically grounded interaction animation is an important step toward learning humanoid-object behaviors from the open visual world.

Acknowledgments and Disclosure of Funding

This work was done during the first author's Tencent Games Motus internship, and supported in part by the Research Grant Council of the Hong Kong SAR under Theme-based Research Scheme, grant no. T22-606/23R.

References

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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15935–15946, 2022.

Yukang Cao, Jiahao Lu, Zhisheng Huang, Zhuowen Shen, Chengfeng Zhao, Fangzhou Hong, Zhaoxi Chen, Xin Li, Wenping Wang, Yuan Liu, et al. Reconstructing 4d spatial intelligence: A survey. *arXiv preprint arXiv:2507.21045*, 2025.

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

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 Computer Vision and Pattern Recognition Conference*, pages 1783–1794, 2025.

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*, pages 22605–22615, 2025.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954, 2023.

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*, pages 14783–14794, 2023.

Google. Gemini 3.1 Flash Image Preview (Nano Banana 2). Google AI for Developers, 2026. AI image generation model; model ID: gemini-3.1-flash-image-preview. Accessed 27 April 2026.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019.

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*, 132(7):2551–2566, 2024.

Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022.

Ao Li, Jinpeng Liu, Yixuan Zhu, and Yansong Tang. Scorehoi: Physically plausible reconstruction of human-object interaction via score-guided diffusion. *arXiv preprint arXiv:2509.07920*, 2025.

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. *arXiv preprint arXiv:2601.01222*, 2026.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. Easyhoi: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7037–7047, 2025.

Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4world: Feedforward world-centric dense 3d tracking of all pixels. *arXiv preprint arXiv:2603.02573*, 2026.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Hamer: Hand mesh recovery for the egoexo4d hand pose challenge.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

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 (TOG)*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 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 2024 Conference Papers*, pages 1–11, 2024.

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*, pages 2070–2080, 2024.

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

Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. *arXiv preprint arXiv:2410.03441*, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025a.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

Zhenrong Wang, Qi Zheng, Sihan Ma, Maosheng Ye, Yibing Zhan, and Dongjiang Li. End-to-end hoi reconstruction transformer with graph-based encoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27706–27715, 2025c.

Zihan Wang, Jiashun Wang, Jeff Tan, Yiwen Zhao, Jessica Hodgins, Shubham Tulsiani, and Deva Ramanan. Contact-guided real2sim from monocular video with planar scene primitives. In *International Conference on Learning Representations*, pages 115452–115468, 2026.

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*, pages 17868–17879, 2024.

Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. *Advances in Neural Information Processing Systems*, 37:125116–125141, 2024.

Yan Xia, Xiaowei Zhou, Etienne Vouga, Qixing Huang, and Georgios Pavlakos. Reconstructing humans with a biomechanically accurate skeleton. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5355–5365, 2025.

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*, pages 21469–21480, 2025.

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*, pages 125–145. Springer, 2022.

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*, pages 4757–4768, 2023.

Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. Intertrack: Tracking human object interaction without object templates. In *2025 International Conference on 3D Vision (3DV)*, pages 1427–1439. IEEE, 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 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. *arXiv preprint arXiv:2404.07191*, 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 Computer Vision and Pattern Recognition Conference*, pages 12266–12277, 2025.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11097–11106, 2021.

Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, Matt Feiszli, Jitendra Malik, Piotr Dollar, and Kris Kitani. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

Haoyu Zhang, Wei Zhai, Yuhang Yang, Yang Cao, and Zheng-Jun Zha. End-to-end spatial-temporal transformer for real-time 4d hoi reconstruction. *arXiv preprint arXiv:2603.14435*, 2026.

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

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*, pages 34–51. Springer, 2020.

Chengfeng Zhao, Juze Zhang, Jiashen Du, Ziwei Shan, Junye Wang, Jingyi Yu, Jingya Wang, and Lan Xu. I'm hoi: Inertia-aware monocular capture of 3d human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–741, 2024.

Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

This appendix provides additional implementation details, and further discussions that support the main paper.

Appendix A. Technical Details: object scale estimation (A.1), object tracking (A.2), contact optimization (A.3), convex decomposition (A.4), and physics simulation (A.5).

Appendix B. Evaluation Details: metric definitions (B.1), repeated-run evaluation (B.2), and computational requirements (B.3).

Appendix C. Limitations and Future Work: reconstruction limitations, interaction scope (C.1), and future directions (C.2).

A Technical Appendices

A.1 Object Scale Estimator

Monocular depth estimation produces metric-ambiguous depth maps. To resolve the scale ambiguity for object reconstruction, we employ a unified scale estimator that combines silhouette IoU and depth consistency:

$$s^* = \text{argmax}_{s \in [s_{\min}, s_{\max}]} \text{IoU}(M_{\text{pred}}(s), M_{\text{obs}}) \cdot \exp\left(-\frac{1}{2}\left(\frac{\Delta z(s)}{\sigma_{\text{depth}}}\right)^2\right) \tag{5}$$

where s is the candidate scale factor, $M_{\text{pred}}(s)$ is the rendered object silhouette at scale s, M_{obs} is the observed mask from SAM3, $\Delta z(s)$ is the mean depth residual between rendered and observed depth, and $\sigma_{\text{depth}} = 0.05\text{m}$ is the depth tolerance.

We search over a discrete set of scales (default: 13 steps from 0.5 to 1.5) on a subset of clean reference frames (default: 8 frames) selected by quality gates. A self-consistency re-pass is triggered if $|s^* - 1| > 0.25$, re-running FoundationPose registration with the scaled mesh.

A.2 Object Tracker

A.2.1 FoundationPose + Healthy Gated Tracker

We extend FoundationPose Wen et al. (2024) with a health-gated tracking mechanism to improve robustness. The tracker maintains a health score $h_t \in [0, 1]$ for each frame based on:

$$h_t = \alpha_{\text{depth}} \cdot h_t^{\text{depth}} + \alpha_{\text{mask}} \cdot h_t^{\text{mask}} \tag{6}$$

where h_t^{depth} measures depth consistency and h_t^{mask} measures mask IoU between rendered and observed masks.

The tracker operates in three modes:

- **Healthy tracking** ($h_t > \theta_{\text{ok}} = 0.20$): Standard FoundationPose tracking
- **Degraded tracking** ($\theta_{\text{deg}} = 0.15 < h_t \leq \theta_{\text{ok}}$): Increase refinement iterations
- **Re-registration** ($h_t \leq \theta_{\text{deg}}$): Trigger full pose estimation from scratch

A cooldown period (default: 10 frames) prevents rapid re-registration oscillations. Optionally, frozen tracking stretches (low-confidence segments) can be interpolated from neighboring healthy frames.

A.3 Contact Optimization

A.3.1 Four-Stage Optimization

Our contact optimization proceeds in four stages to progressively refine the reconstruction:

Stage 1: Depth Pre-optimization Optimize SMPL translation to align with depth maps while maintaining contact constraints:

$$\mathcal{L}_{\text{depth-pre}} = w_c\mathcal{L}_{\text{contact}} + w_p\mathcal{L}_{\text{pen}} + w_t\mathcal{L}_{\text{temporal}} \tag{7}$$

Iterations: 100, LR: 0.005, weights: $w_c = 50$, $w_p = 20$, $w_t = 3$.

Stage 2: Object Pose Pre-optimization Optimize object 6-DoF poses with contact and temporal constraints:

$$\mathcal{L}_{\text{obj-pre}} = w_c\mathcal{L}_{\text{contact}} + w_p\mathcal{L}_{\text{pen}} + w_{t,\text{trans}}\mathcal{L}^{\text{trans}}_{\text{temporal}} + w_{t,\text{rot}}\mathcal{L}^{\text{rot}}_{\text{temporal}} + w_a\mathcal{L}_{\text{anchor}} \tag{8}$$

Iterations: 150, LR: trans $= 0.003$, rot $= 0.001$, weights: $w_c = 50$, $w_p = 20$, $w_{t,\text{trans}} = 3$, $w_{t,\text{rot}} = 2$, $w_a = 1$.

Stage 3: Joint Optimization Jointly optimize SMPL body pose and object poses:

$$\mathcal{L}_{\text{joint}} = w_c\mathcal{L}_{\text{contact}} + w_p\mathcal{L}_{\text{pen}} + w_{t,\text{smpl}}\mathcal{L}^{\text{smpl}}_{\text{temporal}} + w_{t,\text{obj}}\mathcal{L}^{\text{obj}}_{\text{temporal}} + w_a\mathcal{L}_{\text{anchor}} \tag{9}$$

Iterations: 300, LR: body $= 0.0005$, obj-trans $= 0.003$, obj-rot $= 0.001$.

Stage 4: Hand Refinement Refine hand poses using HaMeR priors:

$$\mathcal{L}_{\text{hand}} = w_c\mathcal{L}_{\text{contact}} + w_p\mathcal{L}_{\text{hand-pen}} + w_{\text{pca}}\mathcal{L}_{\text{hand-pca}} + w_t\mathcal{L}^{\text{hand}}_{\text{temporal}} \tag{10}$$

Iterations: 200, LR: 0.01, weights: $w_{\text{pca}} = 3$, hand PCA components: 12.

A.3.2 Loss Function Details

Contact Loss Bidirectional chamfer distance between contact regions:

$$\mathcal{L}_{\text{contact}} = \frac{1}{|V_h|}\sum_{v\in V_h}\min_{u\in V_o}\|\mathbf{v}-\mathbf{u}\|_2^2 + \frac{1}{|V_o|}\sum_{u\in V_o}\min_{v\in V_h}\|\mathbf{u}-\mathbf{v}\|_2^2 \tag{11}$$

where V_h and V_o are vertices in human and object contact regions.

Penetration Loss Penalize interpenetration using signed distance:

$$\mathcal{L}_{\text{pen}} = \sum_{v\in V_{\text{pen}}}\max(0,-d(v))^2 \tag{12}$$

where $d(v) < 0$ indicates penetration depth.

Temporal Smoothness Acceleration penalty:

$$\mathcal{L}_{\text{temporal}} = \sum_{t=1}^{T-1}\|\theta_t - 2\theta_{t-1} + \theta_{t-2}\|_2^2 \tag{13}$$

A.4 Convex Decomposition

For physics simulation in IsaacGym, both human and object meshes require convex decomposition:

- **Object**: V-HACD with parameters: max convex hulls: 32, max vertices per hull: 64, resolution: 300000
- **Humanoid**: Custom MuJoCo XML articulated body generated from SMPL-X with convex collision geometry

The decomposed collision geometry is used for contact detection and physics simulation, while the original high-resolution mesh is used for rendering.

A.5 Physics Simulation

Our physics simulator uses IsaacGym with PPO reinforcement learning. The humanoid is represented as a custom MuJoCo XML articulated body generated from SMPL-X with β-dependent link dimensions and masses.

Each reconstructed object is modeled as a uniform-density rigid body, with mass and inertia determined by its collision geometry and the prescribed density. Since the evaluated datasets do not provide sequence-specific physical-property annotations, we do not attempt to recover the object's true mass, inertia, friction, or material properties from monocular observations. Instead, we adopt a standardized rigid-body configuration, using identical density and contact parameters across all compared teacher trajectories. This controlled setting isolates the effect of the reconstructed kinematic teachers on physical executability, rather than confounding the comparison with differences in simulator parameters.

A.5.1 Reward Function

$$r_t = w_{\text{pose}} r_{\text{pose}} + w_{\text{vel}} r_{\text{vel}} + w_{\text{contact}} r_{\text{contact}} + w_{\text{obj}} r_{\text{obj}} \tag{14}$$

where:

- $r_{\text{pose}} = \exp(-\|\mathbf{q}_t - \mathbf{q}_t^{\text{ref}}\|_2^2)$: Pose tracking reward
- $r_{\text{vel}} = \exp(-\|\dot{\mathbf{q}}_t - \dot{\mathbf{q}}_t^{\text{ref}}\|_2^2)$: Velocity tracking reward
- $r_{\text{contact}} = \mathbb{1}[\text{contact matches reference}]$: Contact reward (warmup: 3 steps)
- $r_{\text{obj}} = \exp(-\|\mathbf{p}_t^{\text{obj}} - \mathbf{p}_t^{\text{obj,ref}}\|_2^2)$: Object tracking reward

Weights: $w_{\text{pose}} = 0.5$, $w_{\text{vel}} = 0.1$, $w_{\text{contact}} = 0.2$, $w_{\text{obj}} = 0.2$.

A.5.2 Training Details

- **Algorithm**: PPO with GAE ($\lambda = 0.95$, $\gamma = 0.95$)
- **Network**: Actor-critic with separate networks
  - Actor: 5-layer MLP (2048-2048-1024-1024-512 hidden units, GELU activation)
  - Critic: 5-layer MLP (2048-2048-1024-1024-512 hidden units, GELU activation)
- **Optimization**: Adam optimizer, actor LR: 5×10^{-6}, critic LR: 1×10^{-4}
- **Training schedule**: Horizon: 8, batch size: 256, optimization epochs: 5
- **Parallel envs**: 512 (configurable)
- **Episode length**: 150 frames (configurable)
- **Max training epochs**: 1,000,000 (varies by sequence complexity)

B Evaluation Details

B.1 Evaluation Metrics Definitions

Chamfer Distance (CD). For meshes $\mathcal{M}_{\text{pred}}$ and $\mathcal{M}_{\text{gt}}$ with vertex sets V_{pred} and V_{gt}:

$$\text{CD}(\mathcal{M}_{\text{pred}}, \mathcal{M}_{\text{gt}}) = \frac{1}{|V_{\text{pred}}|} \sum_{v \in V_{\text{pred}}} \min_{u \in V_{\text{gt}}} \|v - u\|_2 + \frac{1}{|V_{\text{gt}}|} \sum_{u \in V_{\text{gt}}} \min_{v \in V_{\text{pred}}} \|u - v\|_2 \tag{15}$$

Acceleration Error (Acc). For joint positions $\mathbf{j}_t$ at frame t:

$$\text{Acc} = \frac{1}{T-2} \sum_{t=1}^{T-2} \left\| (\mathbf{j}_{t+1} - 2\mathbf{j}_t + \mathbf{j}_{t-1}) - (\mathbf{j}_{t+1}^{\text{gt}} - 2\mathbf{j}_t^{\text{gt}} + \mathbf{j}_{t-1}^{\text{gt}}) \right\|_2 \tag{16}$$

Penetration (Pen). In our setting, we measure human-body vertices penetrating the object:

$$\mathrm{Pen} = \frac{1}{T|V_{\mathrm{human}}|} \sum_{t=1}^{T} \sum_{v \in V_{\mathrm{human}}^t} \max\big(0, -d_{\mathrm{obj}}^t(v)\big), \tag{17}$$

where V_{human}^t are the human mesh vertices at frame (t), and $d_{\mathrm{obj}}^t(v)$ is the signed distance from human vertex (v) to the posed object mesh at frame (t). Negative values indicate that the human vertex lies inside the object. The result is reported in centimeters.

We evaluate interaction quality at two complementary levels: geometric contact recovery before physics projection and interaction persistence after physical rollout.

Contact Recall (C-Recall). Contact recall measures how accurately the reconstructed human-object configuration recovers ground-truth contacts. Let $\mathcal{C}_t^{\mathrm{gt}}$ denote the set of ground-truth contact samples at frame t, and let $d_{\mathrm{pred}}(v,t)$ denote the distance from a ground-truth contact sample to the corresponding predicted contact surface. We compute frame-weighted recall at a threshold of $\delta = 1.5$ cm:

$$\text{C-Recall} = \frac{\sum_{t=1}^{T} |\mathcal{C}_t^{\mathrm{gt}}| \left(\frac{1}{|\mathcal{C}_t^{\mathrm{gt}}|} \sum_{v \in \mathcal{C}_t^{\mathrm{gt}}} \mathbb{1}[d_{\mathrm{pred}}(v,t) < \delta] \right)}{\sum_{t=1}^{T} |\mathcal{C}_t^{\mathrm{gt}}|}. \tag{18}$$

Contact recall is evaluated on the reconstructed kinematic trajectory before physics projection.

Rollout Coverage (RC). Rollout coverage measures how long the reconstructed interaction can be sustained under physical simulation. Let L_{valid} denote the length of the longest continuous valid physical rollout and T the reference sequence length:

$$\mathrm{RC} = \frac{L_{\mathrm{valid}}}{T} \times 100\%. \tag{19}$$

Contact Retention (CR). Contact retention measures the persistence of intended human-object contact after physics projection. Let $\mathcal{T}_{\mathrm{contact}}$ denote the frames with intended contact in the reference interaction, and let $c_t^{\mathrm{sim}} \in \{0,1\}$ indicate whether the intended contact is maintained in simulation:

$$\mathrm{CR} = \frac{\sum_{t \in \mathcal{T}_{\mathrm{contact}}} c_t^{\mathrm{sim}}}{|\mathcal{T}_{\mathrm{contact}}|} \times 100\%. \tag{20}$$

Evaluation Protocol. For physics-based evaluation, all compared teacher trajectories are evaluated using identical simulator configurations, policy architectures, reward functions, training budgets, and termination criteria. Thus, the only experimental variable is the reconstructed kinematic teacher trajectory.

B.2 Statistics

To improve the reliability of our empirical evaluation, we repeat each experiment over 10 independent runs and report the averaged results in the main tables. The repeated runs use the same input videos, dataset split, evaluation protocol, and metric definitions, while accounting for stochasticity introduced by optimization, initialization, tracking recovery, and sequence-specific physics simulation.

This repeated-run protocol is used for the proposed method and its ablation variants. Since several baseline results are taken from prior published evaluations under the standard BEHAVE protocol, direct paired statistical tests against all baselines are not always possible. We therefore use the 10-run averages to assess the stability of our own pipeline and ablations, and report aggregate metrics following the evaluation convention of prior monocular 4D HOI reconstruction work. We also report the $\pm$ 1SD error bar below.

Method	CD-h	CD-o	CD-c	Acc-h	Acc-o	Pen.
Ours	7.11±0.01	15.68±0.11	10.74±0.09	0.52±0.01	0.59±0.02	0.013±0.001

Table 4: Quantitative experiments on BEHAVE test set with $\pm$ 1SD confidence interval.

B.3 Computational Requirements

We measure the computational cost of the full framework on a single NVIDIA H200 GPU workstation with 1.8TB system RAM. On BEHAVE, processing a 1500-frame monocular RGB video takes approximately 47 minutes from input video to the final simulation-ready human-object interaction trajectory.

The average runtime is therefore approximately 1.88 seconds per frame for a 1500-frame sequence. Actual runtime may vary depending on sequence length. In particular, heavily occluded sequences may require additional tracking recovery or contact refinement, while shorter and visually cleaner sequences are processed more efficiently.

C Limitations and Future Work

C.1 Limitations

Despite substantially improving interaction consistency and physical executability, *HA-HOI* has several limitations.

Modular Reconstruction and Perceptual Uncertainty. Our modular design is motivated by the limited availability of diverse, paired 4D HOI supervision. Rather than learning the entire reconstruction problem from a restricted HOI distribution, *HA-HOI* leverages broadly pretrained geometric and motion priors while concentrating interaction-specific reasoning in human anchoring and contact-guided optimization.

However, as our reconstruction stage remains largely vision-based, prolonged occlusion, inaccurate object reconstruction, tracking drift, or incorrect contact predictions may compromise the recovered interaction. Although our progressive contact optimization jointly refines object pose, hand articulation, and selected body joints under geometric and temporal constraints, it cannot fully resolve severe visual ambiguity or consistently recover from erroneous upstream estimates.

Scope of human-anchored reconstruction. Our human-first, object-follow formulation primarily targets human-driven manipulation. However, interactions dominated by external object motion, unexpected resistance, or strong environmental constraints may require more explicitly bidirectional reasoning about human and object trajectories.

C.2 Future Work

A natural next step is to transfer reconstructed interactions to real humanoid robots. One possible pathway is to retarget the recovered trajectories to a target robot embodiment, use them as teacher demonstrations for policy learning in simulation, and subsequently adapt the learned behaviors through system identification, dynamics randomization, and real-world interaction feedback. Such extensions could help bridge the gap between monocular HOI reconstruction and scalable learning of physically grounded humanoid-object behavior.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's scope as monocular 4D HOI reconstruction for simulation-ready human-object interaction, and the claimed contributions are reflected in the proposed human-first, object-follow reconstruction, contact refinement, and physics-based projection pipeline. The experimental claims are limited to reconstruction quality, contact consistency, temporal stability, and simulation readiness under the evaluated BEHAVE and in-the-wild settings.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations in the Discussion / Limitations section, including the reliance on vision-based reconstruction, possible failure under heavy human or hand occlusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present theoretical theorems or formal proof-based results. The equations describe the reconstruction objectives, contact losses, scale estimation, and simulation reward used in the proposed empirical pipeline.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies the input setting, dataset, evaluation split, alignment protocol, metrics, reconstruction stages, contact optimization objectives, ablation settings, and physics simulation design. Additional implementation details are provided in the appendices, including object scale estimation, tracking, contact optimization, convex decomposition, PPO state/action space, rewards, and training details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The experiments use publicly available datasets and existing models where applicable, but the current submission does not yet include an anonymized public code release with full reproduction scripts. We will provide detailed implementation and data-preparation information in the supplemental material to support reproducibility.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper reports the BEHAVE evaluation setting, official train/test split, monocular input protocol, reconstruction metrics, alignment procedure, and ablation settings. The appendices further describe key hyperparameters for object scale estimation, tracking health gates, four-stage contact optimization, convex decomposition, and PPO-based physics simulation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We repeat each experiment over 10 independent runs and report averaged results in the main tables. The repeated-run protocol is described in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the hardware and runtime used for the complete framework.

   Guidelines:

   - The answer [Yes] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses existing datasets and models for scientific evaluation of 4D HOI reconstruction and does not involve new human-subject experiments, private personal data collection, or deployment in a real-world decision-making system. The paper preserves anonymity and discusses responsible scope and limitations of the proposed method.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses positive impacts such as scalable 3D content creation, AR/VR, embodied AI, and simulation-based learning. It also acknowledges potential negative impacts, including misuse for unauthorized motion reconstruction, privacy-sensitive video analysis, or unsafe downstream embodied-agent deployment if physical plausibility is overestimated.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not introduce or release a high-risk pretrained generative model, scraped dataset, or foundation model. The proposed work is a reconstruction and simulation pipeline.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The current draft credits the main datasets, models, and external methods through citations. The license and terms of use explicitly mentioned and properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or other standalone research asset beyond the proposed method. If code is released, it will be documented separately with usage instructions, dependencies, and reproduction scripts.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not conduct crowdsourcing experiments or new research with human subjects. It evaluates the proposed method using existing datasets and in-the-wild videos.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve new human-subject experiments, participant recruitment, or newly collected participant data. Therefore, IRB approval or equivalent review is not applicable to this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use an LLM as an important, original, or non-standard technical component. LLM is only used at word-level to polish writing.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.