SCAN-ANCHORED HIERARCHICAL OPTIMIZATION FOR 4D MEDICAL LOCO-MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Humanoid robots have the potential to assist with specialized medical manipulation tasks, yet collecting large-scale whole-body demonstrations in this domain requires skilled operators, specialized equipment, and substantial human effort. Video generation models provide a scalable alternative, but their outputs often exhibit unstable object motion, inaccurate medical-object geometry, and implausible physical contact which are unacceptable in medical settings. We present **MedManip**, a pipeline that leverages scanned medical objects as geometric anchors to convert generated interaction videos into physically grounded 4D human-object trajectories. Specifically, MedManip progressively propagates the accurate object geometry through three stages: scan-anchored object trajectory estimation, object-guided whole-body motion refinement, and geometry-aware contact optimization. We first recover a temporally stable object trajectory from the generated video, then use the recovered object motion to correct whole-body reachability and human-object alignment, and finally refine fine-grained contact while keeping the object trajectory fixed. The resulting trajectories are retargeted to a humanoid robot and converted into executable whole-body motions through low-level control. In experiments involving more than 80 objects, over **30%** of the generated sequences satisfy our demonstration validity criteria. The resulting policy achieves a **63.6%** task success rate in physics simulation.

1 INTRODUCTION

Humanoid loco-manipulation requires coordinated whole-body motion to approach, interact with, and move objects in complex environments (Dao et al., 2024). Recent advances in whole-body control have enabled humanoid robots to execute increasingly dynamic and contact-rich behaviors, unlocking their potential for real-world applications ranging from household assistance and industrial manipulation to service robotics and human-robot collaboration (Zhang et al., 2025; Fu et al., 2025).

Despite their success in general scenarios, existing motion-capture-based methods are difficult to extend to specialized medical settings due to the high barrier to data collection. Unlike general-purpose data collection, medical procedures require trained professionals to operate specialized equipment and perform complex tasks, making high-quality interaction data particularly difficult to collect at scale. This raises a central question: how can we scale up medical loco-manipulation data while reducing reliance on costly human demonstrations?

An alternative solution is to utilize video generation models to automatically scale up training data, benefiting from their rich priors and strong generalization capabilities in human motion and human-object interactions. Nevertheless, visually plausible generated videos may not be directly applicable to downstream simulation, as they often suffer from unstable physical and geometric consistency in both human motion and object interactions. More importantly, general-purpose video generation models struggle to faithfully preserve the geometry and appearance of specialized medical objects, further limiting their applicability to medical scenarios.

To address these issues, we propose MedManip, a data pipeline that uses scanned object geometry to recover physically grounded 4D interaction trajectories.

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

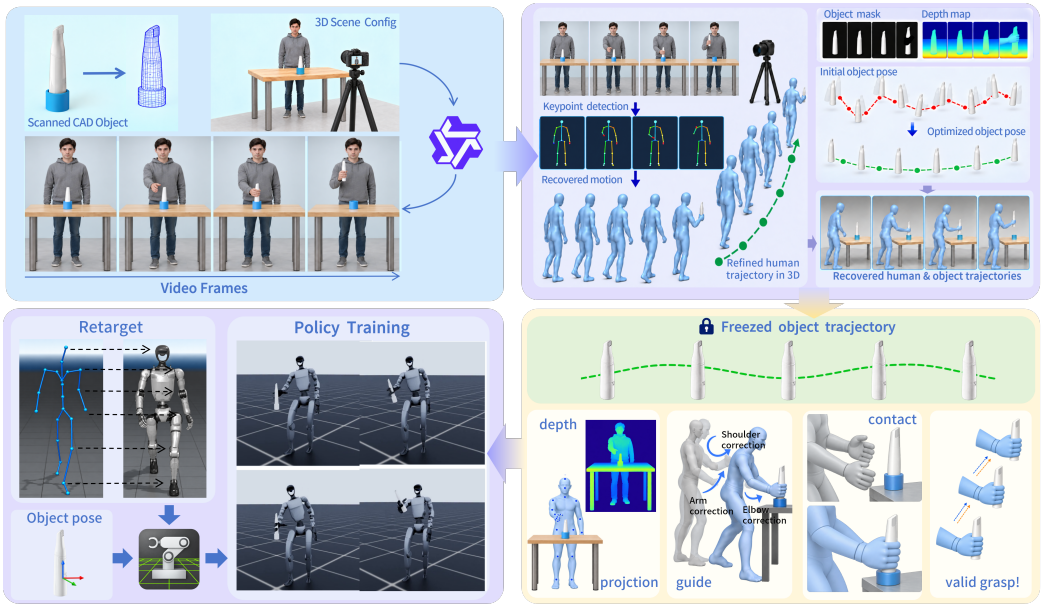

Figure 1: **Overview of MedManip.** Scanned object assets condition video generation and anchor hierarchical recovery of object motion, whole-body coordination, and hand-object contact. The resulting 4D trajectories provide references for humanoid policy training and evaluation in simulation.

Building on the motion priors of GENMO (Li et al., 2025), MedManip uses the scanned object geometry throughout three stages: scan-anchored object trajectory estimation, object-guided whole-body refinement, and geometry-aware contact optimization. Figure 2 illustrates this hierarchy. We first recover a temporally stable object trajectory, then refine whole-body motion using the fixed object trajectory as guidance, and finally enforce stable hand-object contact while reducing penetration.

Through this hierarchical optimization from object motion to whole-body coordination and fine-grained contact, MedManip can convert visually plausible generated videos into simulation-ready manipulation trajectories. Finally, the resulting 4D human-object trajectories are retargeted to a humanoid and used to train a whole-body controller in physics simulation.

We evaluate MedManip on generated interactions involving over **80** objects. Compared with existing generation pipelines, MedManip improves the valid demonstration rate from **17.2%** to **32.6%**, while reducing errors in human-object geometry, object motion, and contact consistency. Physics-based evaluations across diverse loco-manipulation tasks further demonstrate the executability of the reconstructed motions, achieving an overall success rate of **63.6%**.

The main contributions of this work are as follows:

- **A scalable framework for generating physically grounded humanoid loco-manipulation data.** We present MedManip, which converts human-interaction videos synthesized by video-generation models into physically executable whole-body references, thereby reducing the gap between visually plausible generation and robot-ready demonstrations.
- **Temporally consistent human-object geometry and motion recovery.** We combine video depth estimation with explicit 3D object motion cues to reduce depth inconsistency and object trajectory drift in generated videos, providing stable geometric references for subsequent interaction refinement.
- **Contact-aware whole-body interaction refinement.** We infer interaction stages from object motion and progressively refine hand contact, upper-body motion, root position, and foot trajectories. The resulting motion maintains consistent hand-object interaction while preserving feasible whole-body movement.

- **Simulation-based validation of whole-body loco-manipulation.** We evaluate MedManip using geometric and interaction metrics, and validate the executability of the resulting motions with a simulated Unitree G1 across diverse loco-manipulation tasks.

2 Related Works

2.1 Data Generation for Medical Humanoid Manipulation

Large-scale whole-body interaction data are essential for learning humanoid manipulation skills. In daily environments, motion capture and teleoperation systems provide practical ways to collect accurate, robot-executable demonstrations (He et al., 2025b; Ze et al., 2025a;b). These approaches also extend to healthcare applications. Liang et al. (Liang et al., 2026) teleoperate a Unitree G1 to manipulate laparoscopic instruments, while RobOralScan (Lee et al., 2026) learns robotic intraoral scanning in a task-specific simulator. However, medical data collection requires strict safety control, domain expertise, and skilled operators, making physical demonstrations difficult to acquire at scale. Generative video models provide a scalable alternative that reduces reliance on repeated physical data collection. GenMimic (Ni et al., 2025) transfers motions reconstructed from generated human videos to humanoids, while GRAIL (Xie et al., 2026a) and GenHOI (Bi et al., 2026) further extend this paradigm to human-object interactions with object motion and contact. These works establish generative interaction data as a scalable source for humanoid learning, but primarily focus on household interactions. MedManip extends this direction to specialized medical manipulation, where scanned medical tools provide accurate geometric priors for recovering simulation-ready 4D human-object trajectories.

2.2 4D Human-Object Interaction Reconstruction

Recovering 4D human-object interactions from monocular videos requires resolving depth ambiguity while maintaining temporal and contact consistency. MoGe (Wang et al., 2025b) estimates single-image geometry, whereas Video Depth Anything (Chen et al., 2025) incorporates temporal modeling for consistent video depth. D4RT (Zhang et al., 2026), Shape of Motion (Wang et al., 2025a), and Lift4D (Litman et al., 2026) further recover dynamic geometry and 3D motion. Beyond generic 4D reconstruction, interaction-specific methods introduce human-object constraints to recover physically plausible interactions. EgoInfinity (Wang et al., 2026) combines metric calibration with interaction-aware refinement, 4DHOISolver (Wen et al., 2025) uses sparse human-annotated contact correspondences, and GenHOI (Bi et al., 2026) extracts contact constraints from generated videos. HA-HOI (Zhao et al., 2026) and RePHO (Huang et al., 2026) incorporate physics-based simulation to improve the plausibility of interaction. MedManip instead uses scanned object geometry as a reliable geometric anchor and organizes interaction recovery hierarchically, refining the object trajectory before whole-body alignment and fine-grained hand-object contact.

2.3 Humanoid Whole-Body Control and Loco-Manipulation

Humanoid whole-body control provides the foundation for transferring human motion to physically executable robot behaviors. ExBody2 (Ji et al., 2024) combines reinforcement learning with privileged teacher-student distillation to obtain deployable whole-body teleoperation and motion tracking, while ASAP (He et al., 2025a) reduces the Sim2Real dynamics gap through a residual action model from real-world rollouts. BeyondMimic (Liao et al., 2025) combines motion tracking with guided diffusion to compose whole-body skills for downstream tasks, and SONIC (Luo et al., 2026) scales motion data, model capacity, and compute for generalist humanoid control. Beyond motion tracking, humanoid loco-manipulation requires coordinating whole-body motion with object interaction. ResMimic (Zhao et al., 2025) introduces object-conditioned residual learning for whole-body loco-manipulation, while GRAIL (Xie et al., 2026a) transfers generated human–object interactions to humanoid manipulation through physics-based control. We adopt SONIC (Luo et al., 2026) for whole-body control and use the interactions refined by MedManip as reference motions for humanoid loco-manipulation.

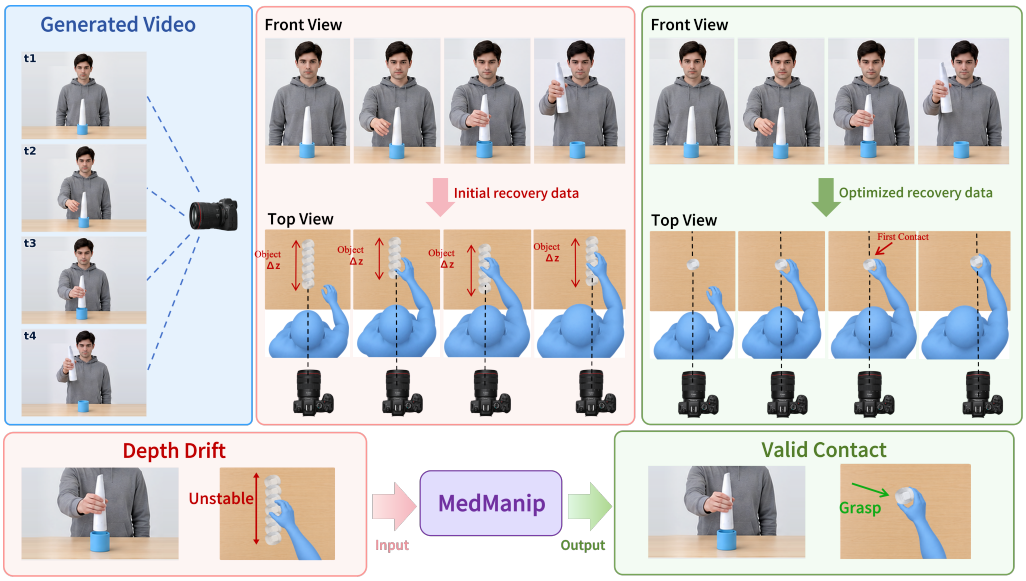

Figure 2: **Hierarchical interaction optimization.** Object trajectory refinement addresses depth drift; the fixed object trajectory then guides whole-body alignment and geometry-aware hand-object contact.

3 METHOD

Given a scanned medical object asset $\mathcal{M}^O$ and a target manipulation task, our goal is to generate physically grounded whole-body motion that can be used for humanoid policy learning. As shown in Fig. 1, MedManip consists of three stages. We first generate human-object interaction videos in a scene configured with the scanned object. We then recover 4D human-object trajectories through a scan-anchored hierarchical optimization that proceeds from object motion $\{\Theta_t^O\}_{t=1}^T$ to whole-body coordination and fine-grained contact $\{\Theta_t^H\}_{t=1}^T$. Finally, the recovered trajectories are retargeted to the humanoid $\{q_t^r\}_{t=1}^T$ and used to train a physics-based whole-body controller.

3.1 SCENE-CONDITIONED MEDICAL INTERACTION GENERATION

We first construct a 3D scene in Blender using a scanned target object, a human model, and the target environment. The scanned asset provides the object's metric geometry and initial pose, while the human and camera are placed according to the target task. We render the initial scene using known camera intrinsics and extrinsics and use it to condition a video generation model with a task-specific prompt, producing a human-object interaction video $\mathcal{V} = \{I_t\}_{t=1}^T$. Unlike unconstrained video generation, this setup provides an explicit geometric reference for the manipulated object before generation. We retain the object model throughout subsequent 4D recovery, allowing its known geometry to anchor object trajectory estimation and human-object interaction refinement.

3.2 SCAN-ANCHORED HIERARCHICAL 4D INTERACTION RECOVERY

Generated interaction videos provide rich motion priors but do not directly yield reliable 4D demonstrations. Independent recovery of human and object motion often introduces depth drift, object trajectory errors, and inconsistent hand-object contact. Jointly optimizing all variables can further propagate errors between the human and object trajectories. We therefore organize 4D interaction recovery hierarchically, as illustrated in Fig. 2. The scanned object serves as the geometric anchor: we first recover and refine its trajectory, then use the stabilized object motion to correct the whole-body trajectory, and finally refine fine-grained hand-object contact while keeping the object trajectory fixed. Figure 3 shows the progressive effect of these stages on the reconstructed interaction.

3.2.1 Initial Human-Object Trajectory Recovery

Given the generated interaction video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, we first obtain initial human and object trajectories using existing reconstruction models. For each frame t, the human state is represented by $\widehat{\Theta}_t^H = (\widehat{g}_t^H, \widehat{\theta}_t^H, \beta)$, and the object state $\widehat{\Theta}_t^O = (\widehat{R}_t^O, \widehat{t}_t^O)$ by its 6-DoF pose. We initialize human motion using GENMO (Li et al., 2025) and estimate the object pose with FoundationPose (Wen et al., 2024) using the scanned object mesh. These estimates provide a coarse 4D interaction trajectory, but remain sensitive to monocular depth ambiguity, occlusion, and frame-wise tracking errors. We therefore refine them progressively in the following stages.

3.2.2 Scan-Anchored Object Trajectory Refinement

We first refine the object trajectory while keeping the recovered human motion fixed. Since the object geometry is known from scanning, each estimated object pose defines an explicit 3D object in the scene. We optimize its trajectory using complementary constraints from image space, depth, and temporal motion.

Projection consistency. Given the scanned object vertices, we project the transformed mesh into each frame and align it with the observed object region:

$$\mathcal{L}_{\text{proj}} = \frac{1}{T}\sum_t \left\|\mathcal{K}^H(\Theta_t^H) - p_t\right\|_1 + \frac{1}{T}\sum_t \left\|\mathcal{P}(V_t^O) - \mathcal{P}(\widehat{V}_t^O)\right\|_1 \tag{1}$$

where p_t denotes the body and hand keypoints detected from the video, and $\mathcal{K}^H$ projects the SMPL-X joints using the known camera parameters. The first term aligns the optimized human motion with the observed keypoints, while the second preserves the image-plane object alignment provided by the initial FoundationPose estimates, where V_t^O and $\widehat{V}_t^O$ denote the object vertices under the optimized and initial poses, respectively.

Temporally consistent depth. Frame-wise monocular depth estimates often fluctuate across time, leading to unstable object positions along the camera axis. We therefore use VideoDepthAnything (Chen et al., 2025) to estimate temporally consistent video depth. We align the scale- and shift-ambiguous predicted depth to the metric scene using the rendered depth of its known background geometry (Appendix A.1). The resulting object depth provides a coarse constraint on its 3D position:

$$\mathcal{L}_{\text{depth}} = \frac{1}{T}\sum_{t=1}^{T}\left[\text{CD}\left(V_t^{H,\text{vis}}, P_t^H\right) + \text{CD}\left(V_t^{O,\text{vis}}, P_t^O\right)\right], \tag{2}$$

where we segment the human and object regions with SAM3 (Carion et al., 2026) and unproject the masked depth measurements into per-frame point clouds P_t^H and P_t^O, $V_t^{H,\text{vis}}$ and $V_t^{O,\text{vis}}$ denote the visible human and object vertices, and CD represents the bidirectional Chamfer distance.

Temporal motion consistency. Depth alone does not fully constrain object motion, particularly under occlusion and fast manipulation. We further extract 3D motion cues with Lift4D (Litman et al., 2026) and use them to regularize inter-frame object displacement:

$$z_t^{\text{L4D}} = \underset{m\in\mathcal{Q}_t}{\text{median}}\, x_{m,t,z}^{\text{L4D}}, \tag{3}$$

where $\mathcal{Q}_t$ represents the collection of reliable points at frame t, and $x_{m,t,z}^{\text{L4D}}$ indicates the camera-depth coordinate of the m-th reliable trajectory point. Following temporal smoothing, we define the relative trajectories as $\delta\mathbf{z}^O = [z_t^O - z_1^O]_{t=1}^T$ and $\delta\mathbf{z}^{\text{L4D}} = [z_t^{\text{L4D}} - z_1^{\text{L4D}}]_{t=1}^T$. The refined object trajectory is obtained by

$$\mathcal{L}_{\text{motion}} = \rho\left(\delta\mathbf{z}^O - s\delta\mathbf{z}^{\text{L4D}}\right) + \alpha_v\rho\left(\Delta\delta\mathbf{z}^O - s\Delta\delta\mathbf{z}^{\text{L4D}}\right) + \alpha_a\left\|\Delta^2\delta\mathbf{z}^O\right\|_1, \tag{4}$$

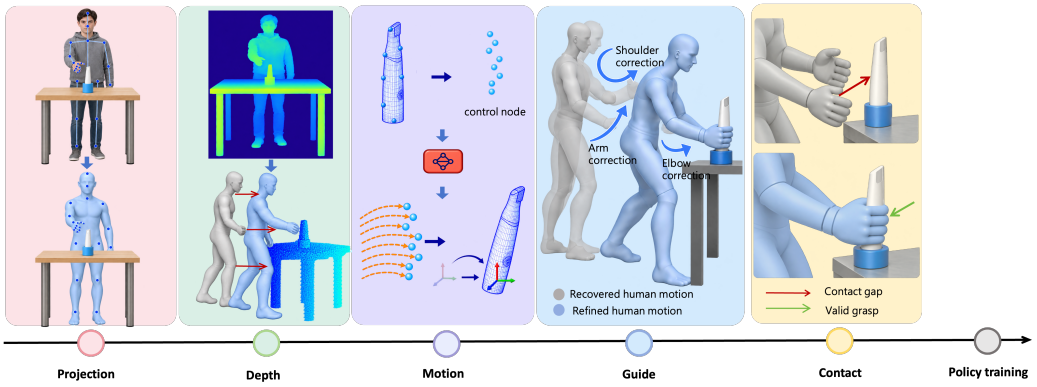

Figure 3: **Progressive interaction refinement.** The initial reconstruction is refined successively for object geometry and motion, whole-body alignment, and hand-object contact.

where s scales the Lift4D trajectory to match the metric scale of the scene, ρ is a robust penalty function, and Δ represents the temporal difference operator. After this stage, we freeze the object trajectory and use it as the geometric reference for subsequent whole-body and contact refinement.

3.2.3 Object-Guided Whole-Body Refinement

A stable object trajectory alone does not guarantee that the reconstructed human can reach the object. We therefore keep the refined object trajectory fixed and adjust human motion to improve their alignment. To identify the transition from approach to manipulation, we analyze temporal changes in the object mask and its image-plane displacement rather than relying solely on individual visual contact cues. We then derive a target palm position from the refined object geometry and propagate the required adjustment through the arm, shoulder, root, and foot trajectory. This improves reachability while retaining the recovered whole-body motion as a prior.

$$\mathcal{L}_{\text{guide}} = \frac{1}{|\mathcal{T}_g|} \sum_{t \in \mathcal{T}_g} \left[\|\text{FK}_h(\mathbf{x}_t) - \mathbf{q}_t^o\|_2^2 + \lambda_{\text{prior}} \|\mathbf{x}_t - \widehat{\mathbf{x}}_t\|_2^2 \right]. \tag{5}$$

where $\widehat{\mathbf{x}}_t$ and $\mathbf{x}_t$ denote the initial and refined human motion parameters at frame t, respectively; $\text{FK}_h(\mathbf{x}_t)$ denotes the palm position obtained through forward kinematics; $\mathbf{q}_t^o$ is the palm guidance target derived from the frozen object pose; $\mathcal{T}_g$ denotes the set of guided frames; and λ_{prior} controls the deviation from the initial human motion. The first term guides the palm toward the object surface, while the second preserves the initial human motion prior. During optimization, we adjust the torso, clavicle, shoulder, elbow, and wrist joints to correct the reaching motion.

3.2.4 Geometry-Aware Contact Refinement

After whole-body alignment, residual errors may still leave a small gap between the hand and object or introduce surface penetration. We therefore refine hand-object contact against the scanned object surface by minimizing the contact distance while penalizing interpenetration:

$$\mathcal{L}_{\text{contact}} = \frac{1}{|\mathcal{T}_c|} \sum_{t \in \mathcal{T}_c} \left[\rho \left(d_{\text{surf}} \left(V_t^{H,\text{cont}}, V_t^O \right) - d^\star \right) + \lambda_{\text{pen}} \mathcal{P} \left(V_t^H, V_t^O \right) \right], \tag{6}$$

where $V_t^{H,\text{cont}}$ denotes the vertices of the hand contact region, V_t^H and V_t^O denote the human and object vertices, respectively, $\mathcal{T}_c$ is the contact interval, d_{surf} measures the hand-to-object surface distance, $d^\star$ is the target contact distance, and $\mathcal{P}$ penalizes hand-object interpenetration. A small nonzero $d^\star$ encourages contact without pulling the hand into the object. During this stage, we keep the refined object trajectory fixed and locally optimize the human pose to maintain stable hand-object contact throughout the interaction.

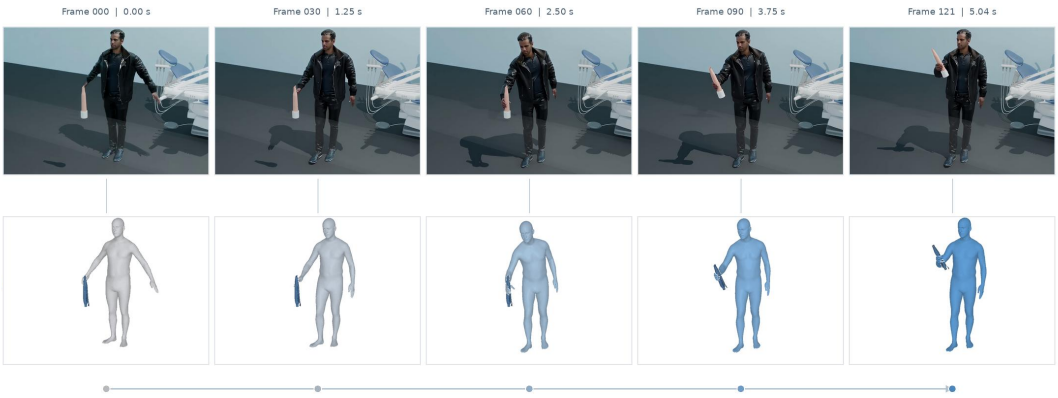

Figure 4: **Camera-view comparison.** Generated video frames and reconstructed human-object motion viewed from the input camera.

Hierarchical optimization. Rather than jointly optimizing all objectives, MedManip follows an object-to-human-to-contact hierarchy. We first optimize the object trajectory using

$$\mathcal{L} = \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{motion}}\mathcal{L}_{\text{motion}} \tag{7}$$

The refined object trajectory is then fixed and used to guide whole-body motion with $\mathcal{L}_{\text{guide}}$, followed by fine-grained hand-object refinement with $\mathcal{L}_{\text{contact}}$.

3.3 Humanoid Loco-Manipulation Policy Learning

The refined human-object trajectories are retargeted to the humanoid morphology to obtain whole-body reference motions. We then train a low-level control policy to reproduce the reference human motion while following the corresponding object trajectory. The policy receives robot proprioception, object state, and reference motion and outputs joint-level actions for whole-body control. We train and evaluate the resulting controller on a simulated Unitree G1, using the corresponding object geometry to preserve the reference established during interaction recovery. Additional details of contact-label generation and policy rewards are given in Appendices A.2 and A.3.

4 Experiments

4.1 Human-Object Interaction Reconstruction Quality

Experimental setup. We compare MedManip with GRAIL (Xie et al., 2026a), an asset-conditioned video generation and 4D HOI reconstruction pipeline; CARI4D (Xie et al., 2026b), a category-agnostic method for reconstructing 4D HOI from monocular videos; and DAViD (Kim et al., 2025), a video-diffusion-based method for generating 4D HOI data. Our evaluation covers more than 80 objects, including everyday objects from RoboCasa Nasiriany et al. (2024) and 3D-scanned medical tools. This combination covers diverse object geometries and interaction patterns; the two scene configurations are described in Appendix A.1.

Evaluation metrics. We evaluate interaction geometry, temporal smoothness, and physical executability. $Contact$ measures the distance between the human contact surface and the object surface during contact frames, while $Pen.$ measures the percentage of human mesh vertices penetrating the object. $HumanSmo.$ and $ObjSmo.$ measure the second-order temporal differences of human and object trajectories, respectively.

We further evaluate physical executability by tracking each reconstructed HOI sequence in physics simulation using InterMimic Xu et al. (2025). $BodyDev.$ measures the mean deviation of simulated human joints from their references, while $ObjDev.$ measures the mean deviation of the simulated object surface. Following GRAIL Xie et al. (2026a), the InterMimic tracking SR is the percentage of evaluated frames for which both deviations, normalized by the object's maximum dimension, are below 0.25. This frame-level metric is distinct from the episode-level policy success rate reported in

Table 1: **Comparison of HOI Generation.** Geometric quality, motion smoothness, physical executability, and reconstruction usability. $Rate$ is the percentage of generated sequences passing the automatic quality and task-motion checks.

Methods	Geometric		Smoothness		Physical Executability			Recon. Usability
	Contact ↓	Pen. ↓	Human Smo. ↓	Obj Smo. ↓	SR (%) ↑	Body Dev. ↓	Obj Dev. ↓	Rate (%) ↑
DAViD Kim et al. (2025)	0.7798	0.7749%	0.0016	0.1596	0.67	0.8212	1.3003	11.4
CARI4D Xie et al. (2026b)	0.4132	1.3847%	0.0049	0.0029	0.63	0.7580	1.0241	8.7
GRAIL Xie et al. (2026a)	0.1107	0.2825%	0.0016	**0.0012**	1.35	0.6131	0.2438	17.2
Ours	**0.02715**	**0.1220%**	**0.0014**	0.0015	**2.99**	**0.5809**	**0.1497**	**32.6**

Table 2: **Tracking Performance on Representative Tasks in Simulation.**

Task	ObjPos (m) ↓	MPJPE-L (mm) ↓
Apple	0.1245	33.31
Avocado	0.1423	45.51
Broccoli	0.2482	53.42
Milk	0.2531	35.98
Oral scanner	0.2312	26.78

Sec. 4.2. $Rate$ is the reconstruction usability rate: the percentage of generated sequences that pass our automatic quality and motion screening. A sequence is retained only when its keypoint-tracking and hand–object contact errors, camera translation, and object and overall mask misalignment remain within their prescribed tolerances. It must also satisfy the motion criteria for the task; in particular, pick-up sequences with no meaningful object motion are rejected. These checks are applied to the reconstructed trajectory before physics-based tracking, so $Rate$ and SR measure distinct properties.

Comparison of reconstruction quality. Figures 4 and 6 provide complementary camera-view and top-view comparisons of the reconstructed interactions. The former shows their alignment with the generated video, while the latter makes spatial trajectory differences easier to inspect. As shown in Table 1, MedManip achieves the lowest contact distance, penetration ratio, and human motion smoothness error. Its reconstruction usability rate rises from 17.2% for GRAIL to 32.6%. The resulting trajectories also attain the highest frame-level InterMimic tracking success rate and the lowest body and object deviations in this comparison. GRAIL has a slightly lower object smoothness error (0.0012 versus 0.0015), so we do not attribute the gains to smoothing alone. Together, these results indicate that improved interaction geometry makes the recovered trajectories more amenable to physics-based tracking.

4.2 Humanoid Policy Training and Execution in Simulation

We further evaluate whether the reconstructed interactions remain executable after transfer to a humanoid robot in simulation. We retarget the recovered trajectories to a Unitree G1 and train a whole-body tracking policy in Isaac Lab. The object is driven entirely by simulated contact dynamics rather than kinematically following its reference trajectory, requiring the robot to reproduce both the reference body motion and object interaction through physical contact.

We evaluate the object position error (ObjPos), local mean per-joint position error (MPJPE-L), and episode-level tracking success rate (Tracking SR). Following GRAIL Xie et al. (2026a), an episode satisfies the tracking criterion when its mean object position error is below 0.20 m. Across the evaluated episodes, the trained policy achieves a Tracking SR of 63.6%, with an ObjPos of 0.1245 m and an MPJPE-L of 35.11 mm. Table 2 gives the task-level errors: apple and avocado have lower object position errors than broccoli and milk, while the oral-scanner task has the lowest local joint error among those listed. Figure 5 shows the evolution of the training indicators. These results validate that the reconstructed interactions can be retargeted to the G1 and executed through physics-based whole-body control, connecting 4D interaction reconstruction with downstream humanoid policy learning.

Table 3: **Ablation of Reconstruction Constraints.** Geometric quality, motion smoothness, and physical executability after removing individual loss terms.

Method	Geometric		Smoothness		Physical Executability		
	Contact ↓	Pen. ↓	Human Smo. ↓	Obj Smo. ↓	SR (%) ↑	Body Dev. ↓	Obj Dev. ↓
w/o L_{depth}	0.6132	**0.0146%**	0.0046	0.0041	1.71	**0.5769**	0.2348
w/o L_{motion}	0.0951	0.1854%	0.0015	0.0026	1.51	0.5857	0.1591
w/o L_{contact}	0.0306	0.0834%	0.0017	**0.0012**	1.43	0.5998	0.1948
Ours	**0.0271**	0.1220%	**0.0014**	0.0015	**2.99**	0.5809	**0.1497**

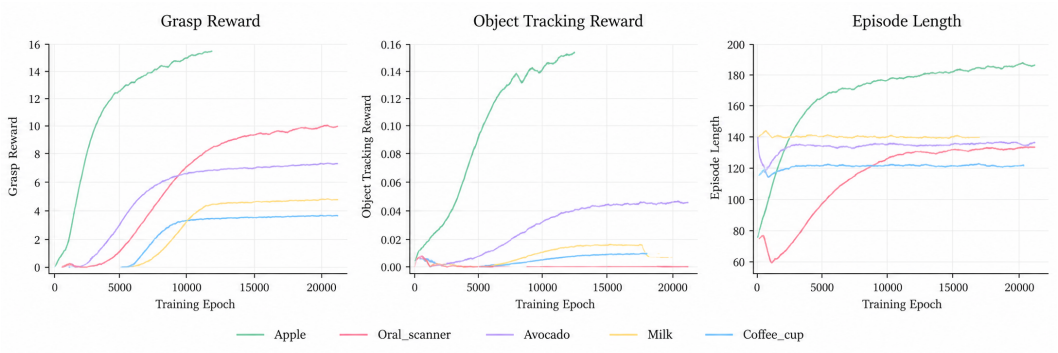

Figure 5: **Policy training curves.** Evolution of the recorded training indicators for the humanoid tracking policy in simulation.

4.3 Ablation Studies on HOI Reconstruction

We ablate the depth, motion, and contact constraints individually to analyze their roles in 4D interaction recovery. All variants are evaluated using the same geometric, temporal, and physics-based metrics as the full model.

Ablation results. As shown in Table 3, removing the depth constraint substantially increases the contact distance from 0.0271 to 0.6132, while increasing the human and object smoothness errors from 0.0014 and 0.0015 to 0.0046 and 0.0041, respectively. The depth constraint is therefore critical for maintaining scene-aligned object geometry.

Without the motion constraint, the object smoothness error increases to 0.0026, and the frame-level InterMimic SR drops from 2.99% to 1.51%, indicating reduced temporal consistency in the object trajectory. Removing the contact constraint further reduces the tracking success rate to 1.43% and increases the object deviation from 0.1497 to 0.1948, highlighting its role in maintaining the recovered interaction under contact dynamics.

Combining all three constraints yields the lowest contact distance, human motion smoothness error, and object deviation, together with the highest physics-based tracking success rate. These results show that depth, motion, and contact provide complementary constraints for recovering physically executable 4D interactions.

5 Conclusion

We presented MedManip, a framework for generating physically grounded 4D interaction data for medical humanoid manipulation. MedManip uses scanned medical objects as geometric anchors and progressively propagates this reliable geometry from object trajectory estimation to whole-body motion refinement and hand–object contact. The resulting trajectories improve interaction geometry and physics-based trackability and can be retargeted to a simulated Unitree G1 for whole-body policy learning and evaluation. MedManip provides a scalable alternative to collecting specialized medical demonstrations from human operators.

Ethics statement

MedManip is developed for research on robotic manipulation and does not involve experiments on patients or the collection of personal medical data. Human motions are generated synthetically and used only as references for robot learning.

Reproducibility statement

We provide the implementation details, optimization objectives, evaluation metrics, and experimental settings needed to reproduce the main results. Code, generated and reconstructed interaction data, and task configurations will be released upon publication.

AI Use Statement

Generative AI models are used as part of the proposed pipeline to generate human-object interaction videos for 4D motion reconstruction. AI tools were also used for language editing and writing assistance. All method design, experiments, analysis, and conclusions were developed and verified by the authors.

References

Zhihai Bi, Qiang Zhang, Guoyang Zhao, Jiahang Cao, Xueyin Luo, Yushan Zhang, Jinglan Xu, Ruoyu Geng, Yulin Li, Andrew F. Luo, and Jun Ma. GenHOI: Contact-aware humanoid-object interaction by imitating generated videos without task-specific training. *arXiv preprint arXiv:2606.12995*, 2026. URL https://arxiv.org/abs/2606.12995. 3

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw. 5

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://arxiv.org/abs/2501.12375. 3, 5, 13

Jeremy Dao, Helei Duan, and Alan Fern. Sim-to-real learning for humanoid box loco-manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)elbow*, pp. 16930–16936. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10610977. 1

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. HumanPlus: Humanoid shadowing and imitation from humans. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 2828–2844, 2025. 1

Tairan He, Jiawei Gao, Wenli Xiao, Yuanhang Zhang, Zi Wang, Jiashun Wang, Zhengyi Luo, Guanqi He, Nikhil Sobanbab, Chaoyi Pan, Zeji Yi, Guannan Qu, Kris Kitani, Jessica Hodgins, Linxi Fan, Yuke Zhu, Changliu Liu, and Guanya Shi. ASAP: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. In *Robotics: Science and Systems*, 2025a. URL https://arxiv.org/abs/2502.01143. 3

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris M. Kitani, Changliu Liu, and Guanya Shi. OmniH2O: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 1516–1540, 2025b. 3

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. URL https://dingbang777.github.io/RePHO/. 3

Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. ExBody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024. URL https://arxiv.org/abs/2412.13196. 3

Hyeonwoo Kim, Sangwon Baik, and Hanbyul Joo. David: Modeling dynamic affordance of 3d objects using pre-trained video diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10330–10341, October 2025. 7, 8

Jinhyung Lee, Haeun Yun, Siwon Kim, Gihyun Baek, Sungho Moon, Sehyun Hwang, and Sunghoon Im. RobOralScan: Learning active intraoral scanning for robotic dental reconstruction. *arXiv preprint arXiv:2606.26955*, 2026. URL https://arxiv.org/abs/2606.26955. 3

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*, pp. 11766–11776, 2025. 2, 5

Zekai Liang, Nikita Thareja, Peihan Zhang, Calvin Joyce, Soofiyan Atar, Florian Richter, Garth Jacobsen, Shanglei Liu, Ryan Broderick, and Michael Yip. In vivo feasibility study of humanoid robots in surgery. *Nature*, 657:236–244, 2026. doi: 10.1038/s41586-026-10796-x. URL https://doi.org/10.1038/s41586-026-10796-x. 3

Qiayuan Liao, Takara E. Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C. Karen Liu. BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion. *arXiv preprint arXiv:2508.08241*, 2025. URL https://arxiv.org/abs/2508.08241. 3

Yehonathan Litman, Xiaoxuan Ma, Manan Shah, Nicolas Ugrinovic, Kris Kitani, Fernando De la Torre, and Shubham Tulsiani. Lift4D: Harmonizing single-view 3D estimation for 4D reconstruction in-the-wild. *arXiv preprint arXiv:2606.23688*, 2026. URL https://arxiv.org/abs/2606.23688. 3, 5

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Fernando Castañeda, Sirui Chen, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, Jinhyung Park, David Sami, Zi Wang, Xingye Da, Runyu Ding, Cyrus Hogg, Lina Song, Edy Lim, Eugene Jeong, Tairan He, Haoru Xue, Wenli Xiao, Simon Yuen, Jan Kautz, Yan Chang, Umar Iqbal, Linxi Fan, and Yuke Zhu. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *Science Robotics*, 11(117):eaed4592, 2026. doi: 10.1126/scirobotics.aed4592. URL https://arxiv.org/abs/2511.07820. 3

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. In *Robotics: Science and Systems (RSS)*, 2024. 7

James Ni, Zekai Wang, Wei Lin, Amir Bar, Yann LeCun, Trevor Darrell, Jitendra Malik, and Roei Herzig. From generated human videos to physically plausible robot trajectories. *arXiv preprint arXiv:2512.05094*, 2025. 3

Gaotian Wang, Kejia Ren, Andrew Morgan, Yiting Chen, Howard H. Qian, Podshara Chanrungmaneekul, and Kaiyu Hang. EgoInfinity: A web-scale 4D hand-object interaction data engine for any-view robot retargeting and video-to-action robot learning. *arXiv preprint arXiv:2606.17385*, 2026. URL https://arxiv.org/abs/2606.17385. 3

Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4D reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a. URL https://arxiv.org/abs/2407.13764. 3

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b. URL https://arxiv.org/abs/2410.19115. 3

Boran Wen, Ye Lu, Keyan Wan, Sirui Wang, Jiahong Zhou, Junxuan Liang, Xinpeng Liu, Bang Xiao, Dingbang Huang, Ruiyang Liu, and Yong-Lu Li. Efficient and scalable monocular human-object interaction motion reconstruction. *arXiv preprint arXiv:2512.00960*, 2025. URL https://arxiv.org/abs/2512.00960. 3

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*, pp. 17868–17879, 2024. 5

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3D assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026a. URL https://arxiv.org/abs/2606.05160. 3, 7, 8, 13, 14

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)*, pp. 14006–14016, June 2026b. 7, 8

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*, pp. 12266–12277, 2025. 7

Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025a. 3

Yanjie Ze, Siheng Zhao, Weizhuo Wang, Angjoo Kanazawa, Rocky Duan, Pieter Abbeel, Guanya Shi, Jiajun Wu, and C. Karen Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025b. 3

Chong Zhang, Wenli Xiao, Tairan He, and Guanya Shi. WoCoCo: Learning whole-body humanoid control with sequential contacts. In *Proceedings of the 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 455–472, 2025. 1

Chuhan Zhang, Guillaume Le Moing, Skanda Koppula, Ignacio Rocco, Liliane Momeni, Junyu Xie, Shuyang Sun, Rahul Sukthankar, Joëlle K. Barral, Raia Hadsell, Zoubin Ghahramani, Andrew Zisserman, Junlin Zhang, and Mehdi S. M. Sajjadi. Efficiently reconstructing dynamic scenes one D4RT at a time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://arxiv.org/abs/2512.08924. 3

Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning. *arXiv preprint arXiv:2510.05070*, 2025. URL https://arxiv.org/abs/2510.05070. 3

Yubo Zhao, Yujin Chai, Yunao Dong, Chengfeng Zhao, Zijiao Zeng, Yuan Liu, and Chi-Keung Tang. Real2Sim in HOI: Toward physically plausible HOI reconstruction from monocular videos. *arXiv preprint arXiv:2605.14462*, 2026. URL https://arxiv.org/abs/2605.14462. 3

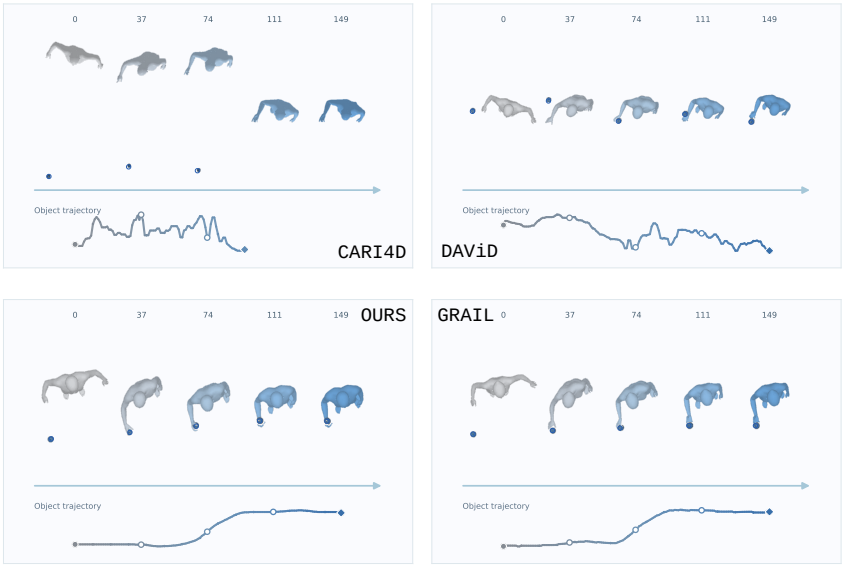

Figure 6: **Top-view comparison.** Human-object motion reconstructed by the compared methods, visualized from above to expose differences in spatial trajectories.

A Supplementary Experimental Details

A.1 Scene Configurations and Video Generation

All scenes used in our experiments are assembled in Blender before video generation. We use two scene configurations. The first is a hospital environment built for interactions with scanned medical objects, including the oral scanner in the representative tracking evaluation. The second is an everyday indoor environment for interactions with household objects. The two settings broaden the object and visual contexts of the evaluation while retaining the same reconstruction and optimization pipeline. Unlike GRAIL's automatically composed candidate scenes (Xie et al., 2026a), both of our environments are explicitly built and configured in Blender.

For each sequence, we place the human character and target object in the selected scene and specify the camera before rendering the initial frame used to condition video generation. Because the scene is constructed before the video is synthesized, the initial positions of both the human and the object, the object mesh and scale, and the camera parameters are known. This provides a common metric coordinate system for human and object initialization. Blender also provides the depth of the static scene geometry, which we use to calibrate the depth predicted by VideoDepthAnything (VDA) (Chen et al., 2025). In our setting, the VDA prediction is temporally consistent but has an unknown metric scale and offset. We fit a single scale a and shift b on static background pixels $\mathcal{B}_1$ of the initial frame:

$$(a^\star, b^\star) = \arg\min_{a,b} \sum_{\mathbf{u}\in\mathcal{B}_1} \rho\big(aD_1^{\mathrm{VDA}}(\mathbf{u}) + b - D_1^{\mathrm{scene}}(\mathbf{u})\big)\,, \qquad D_t^{\mathrm{metric}} = a^\star D_t^{\mathrm{VDA}} + b^\star. \tag{8}$$

Here D_1^{scene} is the rendered metric depth of the known background, $\mathcal{B}_1$ excludes the human and manipulated object, and ρ is a robust penalty. Applying the same calibration to the sequence preserves VDA's temporal depth trend while placing it in the scene's metric frame. The known initial human–object configuration and aligned depth reduce camera-axis ambiguity in subsequent trajectory recovery and provide consistent 3D targets for the object, whole-body, and contact refinement stages.

A.2 Contact Labels from Hand–Object Proximity

GRAIL obtains contact labels by querying a vision-language model on sampled frames (Xie et al., 2026a). We found this unsuitable for our generated views for two reasons. First, the human often faces the camera, so image left and right are reversed relative to the human's anatomical left and

right. A visual-language prediction can therefore select the wrong hand even when the interacting hand is clear in the image. Second, a hand may overlap the object in the image while lying in front of or behind it in 3D; such projections can be misinterpreted as physical contact. Incorrect hand identities or premature contact labels direct contact refinement toward the wrong geometry and can degrade the recovered human and object trajectories.

For the experiments reported here, we instead use a local, frame-wise detector to identify the hand involved in an interaction. It uses the tracked object mask and projected left- and right-hand keypoints; this contact-label path does not call the OpenAI API. The anatomical hand identity follows the known keypoint indexing, and the primary hand is selected from its measured distance to the object rather than a language-model description. The detector still treats image-space proximity as evidence rather than definitive 3D contact, as explained below.

Let M_t^O be the object mask and $\mathcal{K}_{h,t}$ the valid 2D keypoints of hand $h \in \{L, R\}$ at frame t. We compute a distance transform of the object mask and take the closest valid keypoint distance,

$$d_{h,t} = \min_{\mathbf{u} \in \mathcal{K}_{h,t}} D_{M_t^O}(\mathbf{u}), \qquad h \in \{L, R\}, \tag{9}$$

where $D_{M_t^O}(\mathbf{u})$ is the image-plane distance from $\mathbf{u}$ to the nearest object-mask pixel. Invalid or missing hand keypoints are excluded. To avoid declaring contact from a nearby but visibly separate hand, the detector also checks whether the keypoints overlap the object mask. A small disk around each valid keypoint provides a hand support region $M_{h,t}^H$ from which we compute an overlap ratio and mask IoU. The per-frame evidence is

$$E_{h,t} = \mathbb{1}[d_{h,t} \le \tau_d]\, \mathbb{1}\left[r_{h,t} \ge \tau_r \ \vee \ \mathrm{IoU}(M_{h,t}^H, M_t^O) \ge \tau_{\mathrm{IoU}}\right]. \tag{10}$$

Here $r_{h,t}$ is the fraction of valid hand keypoints inside the object mask. The local detector defaults to $\tau_d = 12$ pixels, $\tau_r = 0.50$, and $\tau_{\mathrm{IoU}} = 0.02$, and requires evidence in two consecutive frames before accepting an initial touch candidate. It summarizes active hands in eight-frame intervals. At the selected candidate frame, the primary interacting hand is the one with the smaller $d_{h,t}$; if the consecutive-frame test yields no candidate, the closest valid hand–object observation in the search interval is used as a deterministic fallback.

We distinguish this visual touch candidate from the onset of sustained object motion. The former provides the hand identity and a diagnostic of hand–object proximity, whereas the latter serves as the interaction anchor for trajectory guidance in our reconstruction pipeline. This distinction is useful when the hand approaches or touches the object before the object starts to move, and when occlusion makes a single frame unreliable. Image-plane proximity is used to construct a contact label; the subsequent geometry-aware refinement still checks 3D surface distance and penetration. Thus, a 2D overlap by itself is not treated as proof of physically valid 3D contact.

A.3 Policy Training and Reward Design

The retargeted human and object trajectories provide reference states for policy learning in Isaac Lab. As in GRAIL's object-aware tracking setup (Xie et al., 2026a), the reward combines body-motion tracking, object tracking, grasp-related feedback, and regularization. Our table-top pick-and-place configuration additionally penalizes unintended hand–table contact and excessive approach velocity. The simulated object is dynamic: it must move through contact with the robot rather than following the reference pose kinematically. Consequently, the object reward measures how closely the interaction produced by the policy follows the reference object trajectory.

Table 4 lists the weights explicitly specified in the available table-top training configuration. Other motion-tracking terms, including anchor position, relative body position, and body linear and angular velocity, are inherited from the base reward configuration; their numerical weights are intentionally omitted here because they are not set in this experiment file. The listed weights describe this configuration and should not be read as the weights reported for GRAIL's own policy.

booktabs tabularx

The grasp term is gated by the contact label and, in this configuration, its contact-count component saturates after three detected contacts. The table-contact penalty addresses a failure mode in which the robot reaches toward a table-top object but strikes the support surface instead of grasping the

Table 4: **Reward terms in the table-top pick-and-place configuration.** The table reports weights explicitly set in the experiment configuration. Positive weights reward tracking or grasping; negative weights penalize unsafe contact, abrupt actions, or premature termination.

Category	Configured term	Weight	Role
Motion tracking	`tracking_anchor_ori`	2.5	Tracks the reference root orientation.
	`tracking_relative_body_ori`	5.0	Tracks articulated body orientations relative to the reference.
Object tracking	`object_tracking_reward`	20.0	Encourages the simulated object to follow its reference motion.
Grasping	`grasp_reward`	80.0	Rewards finger–object contact during the labeled grasp phase.
	`grasp_finger_direction`	10.0	Encourages a favorable finger configuration around the object.
Regularization	`finger_primitive_limit`	−10.0	Discourages invalid hand-primitive outputs.
	`meta_action_rate_l2`	−0.1	Penalizes abrupt changes in the policy action.
	`full_latent_rate_l2`	−0.01	Smooths changes in the controller's latent command.
Safety	`hand_table_contact_penalty`	−50.0	Discourages the grasping hand from colliding with the table.
	`approach_velocity_penalty`	−15.0	Reduces excessive hand velocity near the object.
	`is_terminated`	−10.0	Penalizes premature episode termination.

object. We also initialize training episodes from reference states before the labeled contact, sampling the start frame within the first 30 frames. This exposes the policy to the approach phase and the transition into grasping, rather than only to states where the object is already moving. The explicit safety and action penalties complement motion tracking by discouraging a policy from matching the reference body pose through unstable or collision-heavy behavior.