OMNIHOI: DEXTEROUS HAND-OBJECT INTERACTION FROM MONOCULAR HUMAN VIDEO

Anonymous authors
Paper under double-blind review

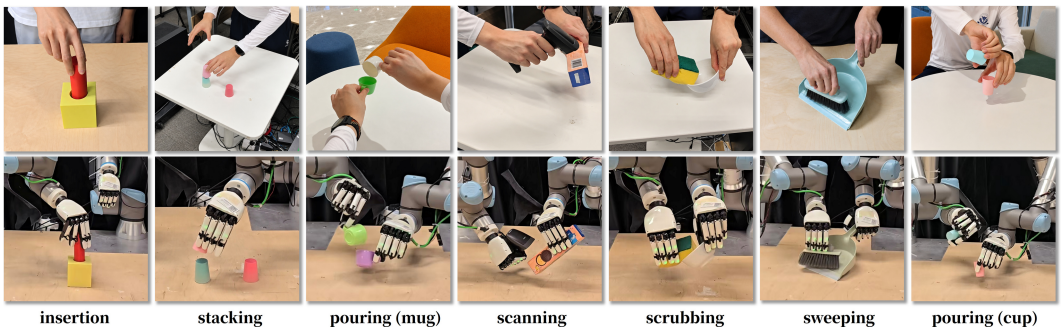


Figure 1: **From human video to robot execution.** Top: monocular RGB videos of human hand-object interaction. Bottom: the trajectories **OmniHOI** produces from them, executed on a pair of XHands.

ABSTRACT

Monocular videos of human manipulation provide abundant dexterous demonstrations, yet reconstructing hand-object interaction from a single view and transferring it to robot hands remain difficult, limiting their direct use for robot execution. Prior methods either require task-specific RL training, limiting scalability, or assume clean motion-capture trajectories and thus cannot operate directly on video. We present **OmniHOI**, a pipeline that turns an RGB video of hand-object interaction into an interaction-faithful trajectory on dexterous hands. The key idea is to enforce physical consistency using the evidence available at each stage: image evidence during reconstruction, contact geometry during retargeting, and dynamics during physics-in-the-loop refinement. Each stage optimizes the corresponding representation directly, correcting errors before they propagate downstream or must be absorbed by a learned policy. Across 150 motion-capture trajectories transferred to each of five dexterous hands with 6 to 22 DoF, we achieve 39–89% success, compared with at most 31% for prior transfer methods. On 60 monocular video clips, we achieve 53% success, compared with 28% for the best prior video-to-robot pipeline. Its trajectories also execute on a real bimanual robot across diverse tasks.

1 INTRODUCTION

Dexterous manipulation demonstrations must capture both hand motion and object interaction. Such data are typically collected through teleoperation or wearable captures (Wang et al., 2024; Hoque et al., 2025), while human manipulation videos are far more abundant. Recent monocular reconstruction models such as MoGe-3 (Kong et al., 2026), SAM 3D (Chen et al., 2025a), and WiLoR (Potamias et al., 2025) can recover scene geometry, object meshes, and hand poses from a single view, making these videos a promising source of robot demonstrations. Turning such video into dexterous robot action requires more than recovering visually plausible motion. Monocular depth ambiguity can confuse occlusion with contact, small geometric errors can turn contact into penetration or separation, and a motion feasible for a human hand may become infeasible on a robot

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.

hand with different morphology and dynamics. These errors must be resolved both when reconstructing the interaction and when transferring it to the robot.

Existing methods make different assumptions about the input and where errors are handled. VideoManip (Chen et al., 2026a) and related methods (Lum et al., 2025; Hsieh et al., 2025; Chen et al., 2026b; Mu et al., 2026) reconstruct hand-object motion from video and then train a task-specific policy, leaving the policy to compensate for reconstruction errors. ManipTrans (Li et al., 2025) and SPIDER (Pan et al., 2025) start from clean motion-capture trajectories and do not address noisy video reconstruction; SPIDER also degrades when the reference motion is noisy. Do as I Do (Paliwal et al., 2026) starts from video without training, but relates the hand and object mainly through centroid alignment and handles contact explicitly only in simulation. Across these pipelines, geometric errors introduced early can survive retargeting and become harder to correct later. A slight finger-object penetration in reconstruction, for example, can turn into a robot link passing through the object after retargeting, leaving the final optimizer with an invalid contact state.

We present **OmniHOI**, a pipeline that reconstructs hand-object interaction from video and transfers it to dexterous robot hands. During reconstruction, dense hand-scene correspondences recover the hand-object geometry, followed by refinement for occlusion, contact, penetration, and finger crossing. During retargeting, the robot hand is optimized to preserve demonstrated contact while removing morphology-induced penetration and crossings. The resulting keyframes are then refined in simulation with CMA-ES optimizer (Hansen & Ostermeier, 2001) to improve object tracking, contact, and grasp stability. Each stage corrects the errors exposed by its own evidence before they propagate to the next stage.

We evaluate **OmniHOI** on reconstruction, motion transfer, and end-to-end video-to-robot execution. It achieves lower hand pose error than prior reconstruction methods while maintaining low hand-object intersection. Across 150 motion-capture trajectories transferred to five dexterous hands with 6 to 22 DoF, we achieve 39–89% success, compared with at most 31% for prior transfer methods. On 60 video clips, we achieve 53% success, compared with 28% for VideoManip and 25% for Do as I Do, and execute the resulting trajectories on a real bimanual robot. Our contributions are:

- A monocular video-to-dexterous-robot pipeline that maintains physical consistency across reconstruction, retargeting, and execution.
- A hand-object reconstruction method that resolves occlusion, contact, penetration, and finger-crossing errors.
- A contact- and dynamics-aware transfer method that produces task-effective motion across diverse dexterous hands.

2 Related Work

Dexterous manipulation from human videos. Human videos have supplied manipulation priors (Nair et al., 2022; Ma et al., 2023; Shaw et al., 2022; Singh et al., 2025; Bahl et al., 2023; Agarwal et al., 2023; Bharadhwaj et al., 2024; Wang et al., 2023), but these require robot data or target grippers, and precise dexterous demonstrations still rely on gloves, AR glasses or headsets (Liu et al., 2025a; Guzey et al., 2025). Closest to ours are pipelines that reconstruct the interaction from video and train a policy on it (Qin et al., 2022), like the systems discussed in Section 1.

Monocular hand-object reconstruction. Hands are recovered robustly in the wild (Pavlakos et al., 2024; Zhang et al., 2025); the held object is predicted by learned models (Hasson et al., 2019; Wu et al., 2024; Chen et al., 2025b) such as iHOI (Ye et al., 2022), lifted with large pretrained models as in EasyHOI (Liu et al., 2025c) and FollowMyHold (Aytekin et al., 2026), or fit per sequence (Fan et al., 2024; Ye et al., 2023; 2024). Fitting-based methods optimize the two together under contact, collision and occlusion constraints (Hasson et al., 2021; Cao et al., 2021; Grady et al., 2021; Jiang et al., 2020; Zhang et al., 2020), whereas robot-learning pipelines such as VideoManip and Do as I Do relate them only through keypoint depth or centroids and leave the contact–occlusion ambiguity open, which **OmniHOI** resolves with pixel-aligned registration and per-finger occlusion-order, contact and crossing corrections.

Human-to-robot motion transfer. Kinematic retargeting, such as WujiHand Retargeting (He & Zhang, 2026), maps human keypoints to robot joints (Handa et al., 2020; Qin et al., 2023; Yin et al.,

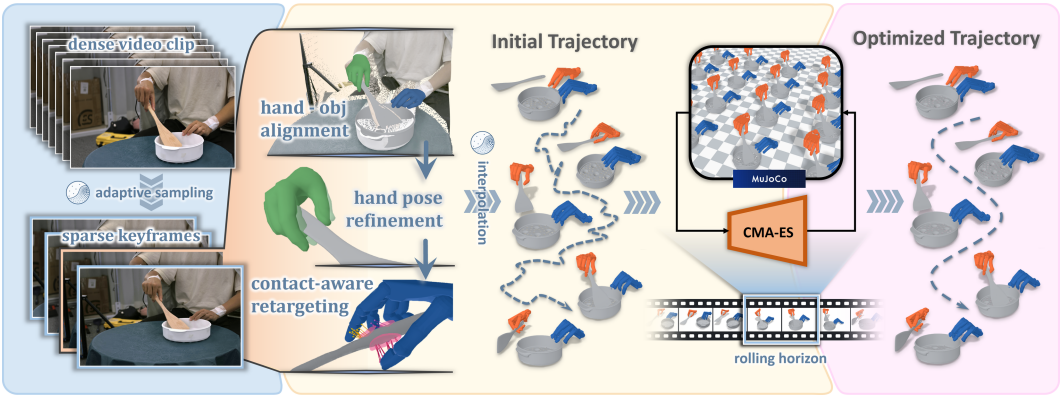


Figure 2: **OmniHOI is built on several key designs.** Adaptive sampling reduces the video clip to keyframes **(left)**. Hand pose in each keyframe is reconstructed, refined and retargeted with contact information preserved **(middle)**. Interpolating between the keyframe configurations then gives an initial robot trajectory. Physics-in-the-loop refinement then rolls the trajectory out in simulation and updates the keyframe configurations with CMA-ES, yielding a task-effective trajectory **(right)**.

2025), and interaction-preserving variants such as OmniRetarget (Yang et al., 2025) keep spatial or contact relations to the object (Ho et al., 2010; Lakshmipathy et al., 2025), but both preserve how the hand looks, not whether it holds the object. Dynamics-aware methods track the reference in simulation through RL or optimization (Chen et al., 2024; Mandi et al., 2025; Liu et al., 2025b; 2024a) or training-free sampling (Williams et al., 2016; Howell et al., 2022; Liu et al., 2010; Ye & Liu, 2012), but, apart from Do as I Do, assume clean motion-capture references. **OmniHOI** resolves contact geometrically before simulation, then refines keyframe offsets with CMA-ES in rolling windows.

3 METHOD

As shown in Figure 2, **OmniHOI** samples keyframes from the video according to hand motion, reconstructs and refines the hand-object interaction in each keyframe (Section 3.1), retargets the refined MANO hands (Romero et al., 2017) to the robot hand under contact guidance, and turns the keyframes into a continuous, task-effective trajectory in simulation (Section 3.2).

3.1 KEYFRAME RECONSTRUCTION

We use SAM 3 (Carion et al., 2025) to segment hands, objects and table, and use MoGe-3 to estimate camera intrinsics and each frame's depth as point map. With a static camera, every point map is aligned to the first frame on table pixels, whose fitted plane later defines the simulation world frame.

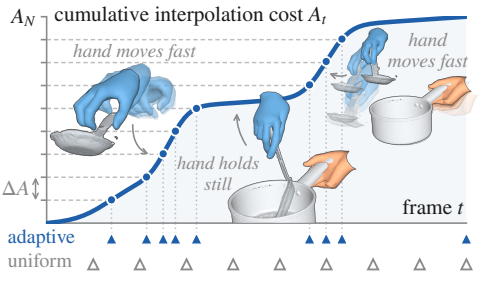


Figure 3: **Keyframes are equal increments of arc length.** Equal steps ΔA of A_t are unequal in time: keyframes crowd where the hand moves fast, unlike uniform sampling of the same budget.

Object-centric hand placement. WiLoR reconstructs the hand mesh up to unknown depth and scale. As mesh and point map are *pixel-aligned*, rasterizing the visible MANO surface pairs each hand-mask pixel with a vertex m_i and a scene point x_i, over which we fit a similarity $(s_h, \mathbf{R}_h, \mathbf{t}_h)$. A first, unweighted fit compares the map with the MANO hand's real size; as monocular metric scale is unreliable, its median over all hands and frames fixes one unit for the clip. In this unit, SAM 3D reconstructs each object once from its most visible frame, and FoundationPose (Wen et al., 2024) registers it per frame from its mask. The final fit

weights each pair by its distance d_i to the object, $w_i = \exp(-\kappa\, d_i/\ell)$ with ℓ the index distal phalanx length, as monocular depth is most reliable near the object.

Adaptive keyframe sampling. Refining every frame is expensive, and at frame scale the coarse poses are dominated by jitter rather than by real motion, so we refine only $r \cdot N$ of the N frames and recover the rest by interpolation. We first accumulate the second-order motion of each hand into

$$A_t = \sum_{u \le t} \left(\omega_{\text{trans}} \left\| \delta^2 \mathbf{t}_u \right\| + \omega_{\text{rot}} \left\| \delta^2 \mathbf{R}_u \right\|_F + \omega_{\text{joint}} \sum_{j=1}^{n_{\text{joints}}} \left\| \delta^2 \boldsymbol{\theta}_{j,u} \right\| \right)^{1/2}, \tag{1}$$

where $\mathbf{t}_u$, $\mathbf{R}_u$ and $\boldsymbol{\theta}_{j,u}$ denote wrist translation, wrist orientation and joint angles at frame u, and $\delta^2(\cdot)$ is the central second-order difference of the smoothed trajectory. Keyframes are placed at equal increments of A_t, so they crowd at reversals and transients and are sparse on steady stretches (Figure 3; details in Appendix B).

Alternating image-contact refinement. The roughly placed hand may still have incorrect occlusion, lost contact or penetration. The two error types are corrected by different evidence and different variables, so we alternate two stages: *image alignment* adjusts the articulation $\boldsymbol{\theta}$ and hand shape $\boldsymbol{\beta}$, *contact-penetration optimization* adjusts $\boldsymbol{\theta}$ together with $\mathbf{T}_h = (\mathbf{R}_h, \mathbf{t}_h)$.

Image-alignment optimization. We render the hand and object depths D_h, $D_{\mathcal{O}}$ and the soft hand silhouette S_h with a differentiable rasterizer (Ravi et al., 2020). For a pixel p in the hand-object overlap Ω, let $\Delta(p) = D_h(p) - D_{\mathcal{O}}(p)$, and $y(p) = +1$ when the hand is visible and $y(p) = -1$ when the object is visible; correct ordering satisfies $y(p)\Delta(p) < 0$ (Figure 4). We penalize violations by

$$\mathcal{L}_{\text{ord}} = \frac{\sum_{p \in \Omega} w(p) \cdot \text{sigmoid}\left(y(p)\Delta(p)/s_{\text{o}}\right)}{\sum_{p \in \Omega} w(p)}, \qquad w(p) = \text{sg}[S_h(p)], \tag{2}$$

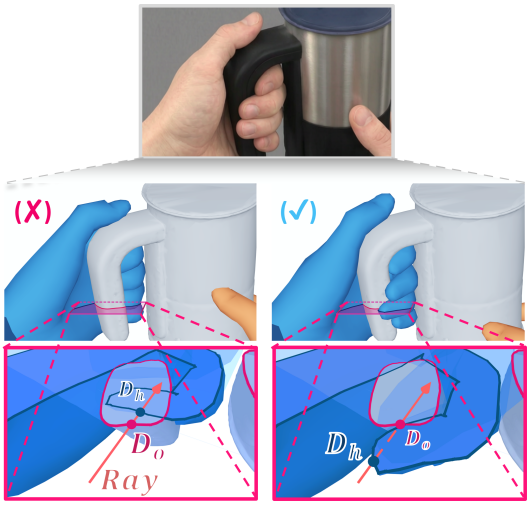
with s_{o} a depth scale. The stop-gradient $\text{sg}[\cdot]$ freezes w during back-propagation, so the hand cannot lower the loss by shrinking its overlap instead of correcting the order. The term is evaluated per finger.

Contact-penetration optimization. This stage attracts each finger's palmar samples P_f to the object surface while repelling its full-surface samples Q_f from the interior. With ϕ the object's signed distance (positive outside) and $\langle \cdot \rangle_P$ an area-weighted mean over P,

$$\begin{aligned} E_f^c &= \left\langle \tfrac{d}{\sigma_c} \log\left(1 + \tfrac{d}{\sigma_c}\right) \right\rangle_{P_f}, \\ E_f^p &= \left\langle \left(\text{ReLU}(-\phi - d_0)/\sigma_p \right)^2 \right\rangle_{Q_f}, \\ \mathcal{L}_{\text{grasp}} &= \frac{1}{5} \sum_{f=1}^{5} \left(\lambda_c E_f^c + \lambda_p\, e^{a_f} E_f^p \right), \end{aligned} \tag{3}$$

where $d = \text{ReLU}(\phi)$, d_0 is a penetration tolerance and a_f the object-surface area enclosed by finger f, an exponential weighting that we found necessary to keep deeply wrapped fingers from sinking into the object. Figure 4 shows the resulting change in finger-object occlusion.

Finger-crossing repair. Signed distance saturates on thin objects: a finger that has *passed through* a plate reads a constant depth and slides inside with no gradient toward the exit (Figure 5). To handle fingers that pass entirely



(a) $\Delta = D_h - D_{\mathcal{O}} > 0$ along the ray $y \cdot \Delta = (+1) \cdot (D_h - D_{\mathcal{O}}) > 0$ (b) $\Delta = D_h - D_{\mathcal{O}} < 0$ along the ray $y \cdot \Delta = (+1) \cdot (D_h - D_{\mathcal{O}}) < 0$

Figure 4: **Occlusion-order loss.** Input image and handle close-ups before (a) and after (b) refinement. A ray through an overlap pixel meets the hand at depth D_h and the object at depth $D_{\mathcal{O}}$; the order is violated in (a) and correct in (b), where the fingers no longer intersect the handle.

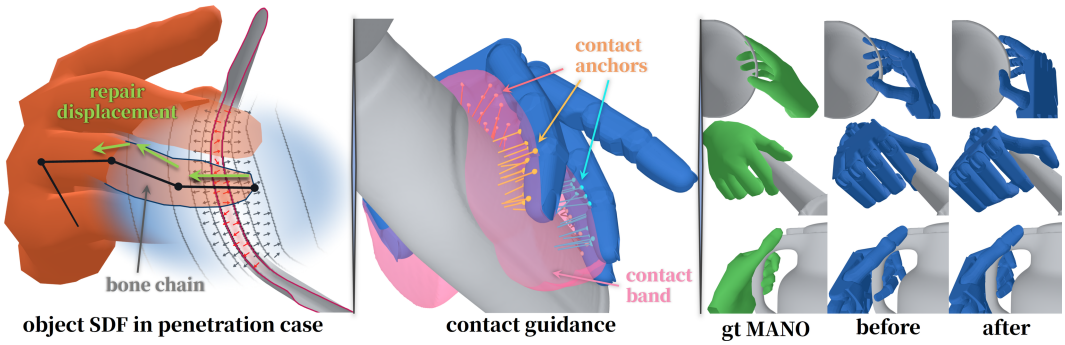


Figure 5: **Contact-aware retargeting. Left:** Arrows denote the gradient directions provided by the SDF. Gradients acting on the same finger link point in opposite directions and cancel each other out, but the repair makes the finger be effectively pushed out of the object along the bone chain. **Middle:** contact guidance: the anchors are sampled on the robot surface within the band around the contact region R_f, which is required to reach. **Right:** comparisons between ground-truth MANO hand, the retargeting solver's output, and the result after repair. With repair, the robot finger remains penetration-free even when the noisy ground-truth MANO annotation already penetrates the object.

through thin objects, we detect surface crossings using parity along each finger's keypoint polyline. We then solve for the minimum pose change that removes the crossing residuals, using a damped linearized update with line search. Full procedure is given in Algorithm 1. The two stages alternate for a fixed number of cycles; repair runs after every cycle. Details are in Appendix A.

3.2 Interaction-faithful human-to-robot transfer

Keyframe reconstruction provides hand poses aligned with the object. To turn these poses into executable robot motion, we first retarget them to the robot hand while preserving contact information, then we interpolate the retargeted keyframes at their source timestamps and refine them in simulation to improve grasp stability and object tracking.

Contact-aware retargeting. We initialize the robot configuration q^{init} with the keypoint-vector solver of WujiHand Retargeting. This preserves the overall finger arrangement, but morphological differences cause contact loss and penetration. We detect hand-object contacts by proximity, assign each contacting finger a contact region R_f on the object and a set of surface anchors, and require its anchors to lie in a narrow band around R_f (Figure 5) and prevent penetration at the same time. Similar to Section 3.1, starting from the kinematic retargeting solution, we iteratively refine wrist and joint configurations. Crossing links are repaired first; otherwise, the update jointly removes penetration/enclosure and pulls contact anchors toward their demonstrated contact regions. We take the smallest pose change that lexicographically reduces the number of crossings and then the residual. See Algorithm 2 for pseudocode.

Physics-in-the-loop trajectory refinement. We interpolate the retargeted keyframes $\{q_k\}$ at their source timestamps $\{t_k\}$ into a 60 Hz trajectory by shape-preserving cubic interpolation (PCHIP; Fritsch & Carlson, 1980), then we replay the trajectory in MuJoCo (Todorov et al., 2012) and search for corrections. The variables for optimization are a per-keyframe offset Δq_k, interpolated the same way and added to the base trajectory.

Contact simulation provides no usable gradients, so we optimize the offsets with CMA-ES, which maintains a multivariate Gaussian over them: each iteration draws samples, scores them by rollouts, and shifts and reshapes it toward the lowest-cost ones, so that it gradually moves toward and contracts around a minimum. The search runs in a rolling window: at keyframe k the simulator resumes from the state saved at $k-1$ and rolls out to $\max(t_{k+1}, t_k + 1\,\text{s})$. Each candidate is scored by

$$\begin{aligned}\mathcal{L} = \lambda_{\text{pos}} E_{\text{pos}} + \lambda_{\text{rot}} E_{\text{rot}} + \lambda_{\text{contact}} E_{\text{contact}} + \lambda_{\text{opp}} E_{\text{opp}} \\ + \lambda_{\text{pen}} E_{\text{pen}} + \lambda_{\text{prior}} \|\Delta q_k\|^2 + \lambda_{\text{term}} E_{\text{term}}.\end{aligned} \tag{4}$$

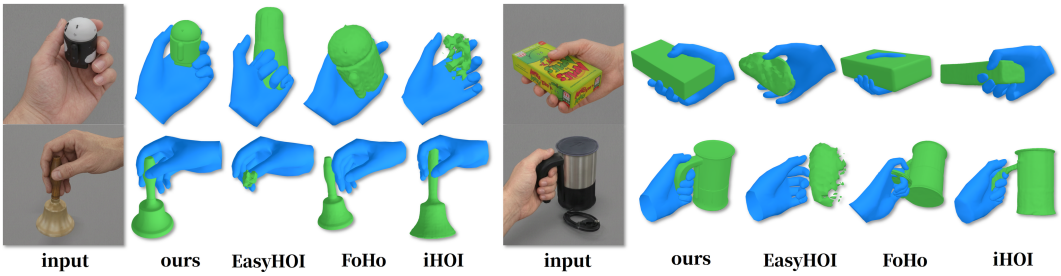


Figure 6: **Qualitative hand-object reconstruction** from a single image. FoHo: FollowMyHold.

E_{pos} and E_{rot} measure the error between simulated object pose and its reconstructed trajectory, and E_{pen} penalizes finger-object penetration. E_{contact} is a penalty on the gap between object and hand links that the retargeted result marks as touching, E_{opp} penalizes the shortfall of the best antagonistic contact pair below a target force, and E_{term} re-applies the pose and contact penalties at the final step of the window, penalizing grasps that look adequate mid-window but have already begun to fail by its end — once Δq_k is frozen, no later keyframe can undo it.

The best offset is then frozen, the simulation advances to t_k, and the search at $k+1$ warm-starts from Δq_k.

4 Experiments

4.1 Hand-object reconstruction

We follow the setup and criteria from prior work, testing our method on three public datasets: ARCTIC (Fan et al., 2023), DexYCB (Chao et al., 2021) and OakInk2 (Zhan et al., 2024), all with 3D hand-object pose and shape annotations. ARCTIC includes humans operating 11 articulated objects, OakInk2 features manipulation tasks with 75 diverse objects, and DexYCB presents grasping of 20 distinct YCB-Video (Xiang et al., 2018) objects. Since prior monocular hand-object reconstruction methods handle only one hand (ours handles both), we compare on 500 randomly sampled images per dataset, each showing one hand in contact with the object.

Table 1: **Reconstruction results.** F5/F10: object F-score at a 5/10 mm surface-distance threshold; CD: object Chamfer distance; MPVPE/MPJPE: mean per-vertex/per-joint error of the hand with dataset ground truth after Procrustes alignment; IV: hand-object intersection volume. Deeper color is better.

Method	Overall						ARCTIC					
	F5 ↑	F10 ↑	CD (cm^2) ↓	MPVPE (mm) ↓	MPJPE (mm) ↓	IV (cm^3) ↓	F5 ↑	F10 ↑	CD (cm^2) ↓	MPVPE (mm) ↓	MPJPE (mm) ↓	IV (cm^3) ↓
EasyHOI	0.2187	0.3890	1.7110	14.63	14.02	23.66	0.1720	0.3067	2.6391	16.20	15.43	27.19
FollowMyHold	0.2007	0.3602	1.7731	8.92	8.48	19.73	0.1698	0.3035	2.6325	10.69	10.10	28.30
iHOI	0.2158	0.3881	1.7384	11.41	10.88	2.90	0.1741	0.3137	2.6672	12.79	12.07	2.35
OmniHOI (ours)	0.3878	0.5831	0.9574	6.05	5.71	2.75	0.2823	0.4568	1.6048	5.36	4.99	2.64
	DexYCB						OakInk2					
	F5 ↑	F10 ↑	CD (cm^2) ↓	MPVPE (mm) ↓	MPJPE (mm) ↓	IV (cm^3) ↓	F5 ↑	F10 ↑	CD (cm^2) ↓	MPVPE (mm) ↓	MPJPE (mm) ↓	IV (cm^3) ↓
EasyHOI	0.2592	0.4672	0.7856	12.93	12.47	28.57	0.2300	0.4013	1.6193	14.62	14.02	14.31
FollowMyHold	0.2127	0.3890	1.0262	5.79	5.60	14.95	0.2239	0.3952	1.5709	10.21	9.67	14.91
iHOI	0.2611	0.4708	0.8182	10.50	10.13	3.31	0.2159	0.3865	1.6404	10.77	10.29	3.10
OmniHOI (ours)	0.4875	0.7159	0.3436	5.03	4.88	4.44	0.4040	0.5883	0.8594	7.94	7.43	1.09

Our pipeline outperforms all baselines (iHOI, FollowMyHold and EasyHOI) on object reconstruction (F5, F10 and CD) and attains substantially lower hand pose errors (MPVPE and MPJPE). Its intersection volume is on par with iHOI (2.75 vs. 2.90 cm^3 overall) and far below those of FollowMyHold and EasyHOI. Figure 6 compares the reconstructions qualitatively.

Table 2: Human-to-robot motion transfer on 150 trajectories per hand across different robot hands. ManipTrans does not support Wuji Hand and SharpaWave. Time is the mean wall-clock minutes to process a trajectory over all hands a method was run on, measured on RTX 4090. It covers per-task RL training and rollout for ManipTrans (excluding imitator pre-training), whole pipeline for SPIDER, and human-to-robot motion transfer for **OmniHOI**. Deeper color is better.

Method	Inspire Hand (6DoF)			XHand (12DoF)			Wuji Hand (20DoF)			SharpaWave (22DoF)			Shadow Hand (20DoF)			
	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Time (min) ↓
ManipTrans	24	9.4	37	22	9.7	33	–	–	–	–	–	–	26	10.7	34	385
SPIDER	26	6.6	43	26	6.2	43	20	8.3	43	19	8.1	54	15	8.5	45	37
OmniRetarget	24	8.4	33	31	3.8	35	27	6.4	43	20	7.3	46	16	8.1	50	0.3
OmniHOI (ours)	39	6.7	26	89	1.8	12	86	1.6	11	89	1.5	10	85	2.4	14	19

4.2 HUMAN-TO-ROBOT MOTION TRANSFER

We compare our method with existing motion transfer approaches to evaluate its ability to convert human hand motion into functionally equivalent dexterous hand action given clean motion-capture data. We select 150 trajectories covering diverse manipulation types, 58 from TACO (Liu et al., 2024b), 62 from OakInk2 and 30 from ARCTIC, of which 60 are single-hand and 90 bimanual (Figure 7), retarget them to five dexterous hands using our method and the baselines (ManipTrans, SPIDER and OmniRetarget), and evaluate manipulation precision, task compliance and transfer efficiency.

Setup. Output trajectories are replayed open-loop in an identical MuJoCo scene. Each scene contains the ground-truth object as a free rigid body with annotated mass, the table, and the remaining static objects. Each hand uses its official model with its joint damping and torque limits; all other simulation settings, including actuator gains, are shared across methods (Appendix D).

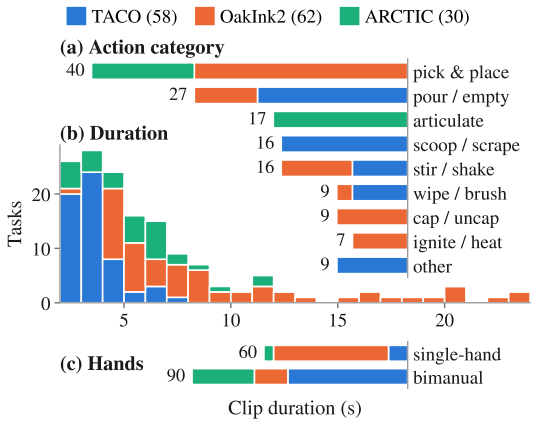


Figure 7: **The 150 transfer tasks** by source dataset. (a) Action category, where "other" is cutting and measuring. (b) Duration of the replayed in-hand clips. (c) Single-hand versus bimanual tasks.

Metrics. For each task we evaluate how far the trajectory of each manipulated object in the replay deviates from its ground-truth trajectory, since the goal is to move the object correctly rather than to precisely imitate the human hand. For each object, we rigidly align the replayed poses $(\mathbf{p}_t, \mathbf{R}_t)$ to the ground truth $(\hat{\mathbf{p}}_t, \hat{\mathbf{R}}_t)$ and take the absolute trajectory error (ATE) $e_{\text{pos}} = \text{RMS}_t \|\mathbf{p}_t - \hat{\mathbf{p}}_t\|$ and $e_{\text{rot}} = \text{RMS}_t \angle(\mathbf{R}_t, \hat{\mathbf{R}}_t)$, with $\angle$ the geodesic angle. A task is successful only if ATE stays below 10 cm and 0.5 rad, and in bimanual tasks both objects must pass. Reported errors are per-task worst-object values, aggregated as the median over the tasks of each hand.

Results. As shown in Table 2, our method achieves the highest success rate on every hand, 85–89% on the four higher-DoF hands where the baselines reach at most 31%, and the lowest object errors except for translation on the Inspire Hand, indicating that the motions it produces have reliable interaction fidelity. The lead holds on all five hands, from 6 to 22 DoF, demonstrating strong cross-embodiment generalization. Success on the 6-DoF Inspire Hand (39%) is nevertheless clearly lower, which we attribute to the limited dexterity of its six motors. In transfer efficiency, **OmniHOI** takes 19 min per trajectory, second only to the purely kinematic OmniRetarget (0.3 min) and below SPIDER (37 min) and ManipTrans (385 min), while succeeding far more often than all three.

4.3 END-TO-END VIDEO-TO-ROBOT TRANSFER

From the 150 videos collected in the previous section, we select the 60 videos with the least hand-object occlusion and the largest hand-object pixel coverage in the frames to evaluate the quality of the outputs produced by the full video-to-action pipeline. They comprise 51 TACO and 9 OakInk2 clips; none comes from ARCTIC, since we avoid articulated objects for reconstruction. We use XHand as robot embodiment for evaluation.

Failure modes. Both baselines estimate the camera with MoGe (Wang et al., 2025a;b). On TACO it reads the 24° telephoto lens as 42–52° and places the whole scene at about half its true distance, so upstream image-based geometry estimation is their largest bottleneck; on OakInk2, whose camera it estimates well, the hand pushes the object aside instead of grasping it, a second bottleneck independent of depth (Appendix F). **OmniHOI** fails on 28 clips, 14 in perception, 13 in execution and 1 in motion planning, most often because the object is reconstructed at the wrong size (12) or rotates in the hand (7).

Table 3: **End-to-end video-to-robot transfer** on the 60 clips, under the ATE criterion of Section 4.2. Succ.Prog.: Progress fraction of the longest prefix that passes the success criterion (mean).

Method	Succ. (%) ↑	Succ.Prog. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓
VideoManip	28	49	8.6	31
Do as I Do	25	48	9.1	31
OmniHOI (ours)	**53**	**71**	**5.7**	**26**

4.4 REAL-WORLD DEPLOYMENT

To demonstrate quality, we execute seven representative self-recorded trajectories on two XHands mounted on UR5e arms, each posing challenges such as bimanual coordination, precise insertion and long-horizon grasp-and-place. We first align the hand-object trajectories produced by our pipeline with the robot workspace in simulation. Since our pipeline operates on floating hands throughout, we then solve IK in simulation to obtain the corresponding arm motions. Finally, we reproduce the simulated scene in the real world and execute the trajectories (Figure 8); further film strips are in Appendix G.

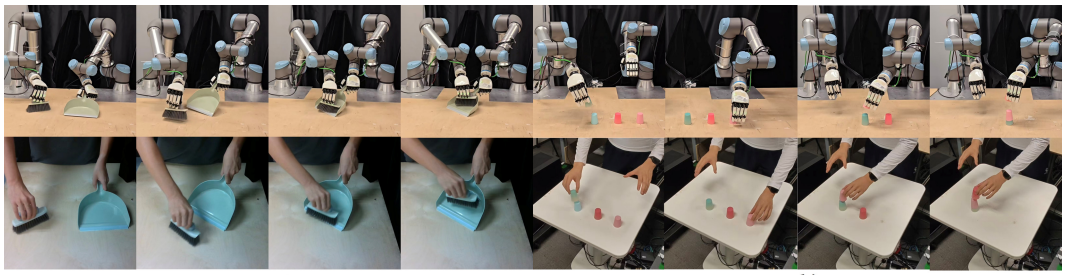


Figure 8: **Real-world deployment.** Two representative tasks, sweeping and stacking, in temporal order from left to right. Top: the robot executing the trajectory. Bottom: input human video.

4.5 ABLATIONS

We ablate reconstruction and transfer separately, each paired over the same frames or tasks; adaptive keyframe sampling is evaluated in Appendix B.

Reconstruction components. Adding the refinement steps of Section 3.1 one at a time to the placed WiLoR hand (Table 4) lowers the intersection volume at every step, by about a third overall (from 4.27 to 2.75 cm^3), while PA-MPVPE and PA-MPJPE rise by only 0.67 mm and remain below every baseline in Table 1. We accept this trade because retargeting inherits its penetration: a robot hand that starts inside the object cannot separate from it in simulation.

Table 4: Reconstruction ablations on the 1500 frames of Table 1: starting from the hand placed as in Section 3.1, the refinement steps of Section 3.1 (image alignment with $\mathcal{L}_{\text{ord}}$, contact-penetration with $\mathcal{L}_{\text{grasp}}$, finger-crossing repair) are added one at a time.

Variant	PA-MPVPE (mm) ↓	PA-MPJPE (mm) ↓	IV (cm^3) ↓
Placement only	**5.37**	**5.03**	4.27
+ image alignment	5.84	5.50	3.61
+ contact-penetration	5.81	5.48	3.03
+ finger-crossing repair	6.05	5.71	**2.75**

Transfer components. We switch contact refinement and physics-in-the-loop refinement on and off on the protocol of Section 4.2 (Table 5); without physics-in-the-loop refinement, the interpolated keyframes are replayed open-loop. Physics-in-the-loop refinement makes the motion executable, raising success from 9–29% to 85–89% on the four higher-DoF hands and from 21–25% to 36–39% on the Inspire Hand. Replayed open-loop, contact refinement lowers success on four hands, mostly by removing false successes: of the 51 tasks it turns into failures, 38 have fingers passing fully through the object in the q^{init} replay (5 after contact refinement), so the interlocked object follows the hand without being held. After physics-in-the-loop refinement it roughly halves the median maximum penetration depth, from 9.17 to 4.73 mm, while success rises by 2–3 points on three hands, is unchanged on the Shadow Hand and drops by 2 on the Wuji Hand (Appendix E).

Table 5: Transfer ablations: the two refinement steps of Section 3.2 switched on and off, on the protocol of Table 2 (150 tasks per hand, from ground-truth MANO, ATE criterion). Deeper color is better.

	Inspire Hand (6DoF)			XHand (12DoF)			Wuji Hand (20DoF)			SharpaWave (22DoF)			Shadow Hand (20DoF)		
Variant	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓
Retargeting (q^{init})	25	8.8	34	29	4.6	38	22	6.8	40	17	7.0	45	12	8.5	53
+ contact refinement	21	8.3	34	23	5.4	44	19	7.5	46	17	7.7	50	9	8.5	59
+ physics-in-the-loop	36	6.8	27	87	1.7	12	88	1.5	10	87	1.8	12	85	2.4	14
+ both	39	6.7	26	89	1.8	12	86	1.6	11	89	1.5	10	85	2.4	14

5 Conclusion

We presented **OmniHOI**, a training-free pipeline that turns a monocular RGB video of hand-object interaction into an interaction-faithful trajectory on a dexterous hand. Instead of absorbing errors into a learned policy, it enforces physical consistency in reconstruction, retargeting and physics-in-the-loop refinement, correcting each error in the stage where it arises. Experiments show lower hand pose error and hand-object intersection in reconstruction, markedly higher success than prior motion-transfer methods on five hands with 6 to 22 DoF, and end-to-end success on videos where prior video-to-robot pipelines fail.

Limitations. **OmniHOI** is currently limited by monocular perception accuracy, especially metric scale estimation. Physics-in-the-loop refinement is also affected by errors in the video-derived object geometry and estimated physical properties, and could benefit from more accurate geometry and dynamics estimation from video. Real-world execution is open-loop, so incorporating feedback control is a natural direction for improving robustness.

AI use statement

In this work, we used generative AI tools to implement methods, to assist with translation, and to clean and reformat datasets. We have not used generative AI tools to help develop theoretical models and conceptual frameworks, to propose or refine hypotheses, to design or provide feedback on research methodology or experiments, to support qualitative and thematic data analysis, or to interpret results. Generating synthetic datasets, formulating mathematical claims, providing critical

ingredients for proving mathematical claims and assisting in the writing of proofs are not applicable to this work.

Additionally, we used generative AI tools to write and edit software code, to suggest experimental parameters, to identify, summarize and analyze existing literature, to format references, and to edit this paper for readability.

We have reviewed all AI-assisted work. Code written or edited with AI assistance was read and tested by the authors, and every number reported in this paper comes from a run the authors executed and inspected themselves. Every reference was checked against its primary source, with author lists and titles verified on the publisher or arXiv page. Experimental parameters suggested by a language model were adopted only after the authors ran them and compared the outcome against the alternatives. Text edited for readability was checked line by line against the authors' intended meaning, and no experimental result reported here was produced or altered by a generative AI tool. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Reproducibility statement

We have made every effort to ensure that the results presented in this paper are reproducible. The experimental setup, including the benchmarks and how evaluation images and trajectories are selected, the robot hands, the simulated replay, the evaluation metrics with their success thresholds, and the baselines, is described in detail in Section 4. We have also provided a full description of **OmniHOI**, covering keyframe reconstruction (Section 3.1), interaction-faithful human-to-robot transfer (Section 3.2), the ablations of its components (Section 4.5), the reconstruction settings (Appendix A), and pseudocode for finger-crossing repair and contact-aware retargeting (Appendix C), to assist others in reproducing our experiments. Additionally, the public hand-object interaction datasets we evaluate on, namely ARCTIC, DexYCB, OakInk2 and TACO, are publicly available, ensuring consistent and reproducible evaluation results. We believe these measures will enable other researchers to reproduce our work and further advance the field.

References

Ananye Agarwal, Shagun Uppal, Kenneth Shaw, and Deepak Pathak. Dexterous functional grasping. In *Conference on Robot Learning (CoRL)*, 2023.

Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow My Hold: Hand-object interaction reconstruction through geometric guidance. In *International Conference on 3D Vision (3DV)*, 2026.

Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Homanga Bharadhwaj, Roozbeh Mottaghi, Abhinav Gupta, and Shubham Tulsiani. Track2Act: Predicting point tracks from internet videos enables generalizable robot manipulation. In *European Conference on Computer Vision (ECCV)*, 2024.

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 (ICCV)*, pp. 12417–12426, 2021.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, et al. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

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

Hongyi Chen, Tony Dong, Tiancheng Wu, Liquan Wang, Yash Jangir, Yaru Niu, Yufei Ye, Homanga Bharadhwaj, Zackory Erickson, and Jeffrey Ichnowski. VideoManip: Dexterous manipulation

policies from RGB human videos via 3D hand-object trajectory reconstruction. *arXiv preprint arXiv:2602.09013*, 2026a.

Kaihan Chen, Yanming Shao, Haifeng Ji, Xiaokang Yang, and Yao Mu. V2P-Manip: Learning dexterous manipulation from monocular human videos. *arXiv preprint arXiv:2606.16436*, 2026b.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, et al. SAM 3D: 3Dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025a.

Yuanpei Chen, Chen Wang, Yaodong Yang, and C. Karen Liu. Object-centric dexterous manipulation from human motion data. In *Conference on Robot Learning (CoRL)*, 2024.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: Monocular hand-held objects reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025b.

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 (CVPR)*, 2023.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 494–504, 2024.

F. N. Fritsch and R. E. Carlson. Monotone piecewise cubic interpolation. *SIAM Journal on Numerical Analysis*, 17(2):238–246, 1980.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. ContactOpt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1471–1481, 2021.

Irmak Guzey, Haozhi Qi, Julen Urain, Changhao Wang, Jessica Yin, Krishna Bodduluri, Mike Lambeta, Lerrel Pinto, Akshara Rai, Jitendra Malik, et al. Dexterity from smart lenses: Multi-fingered robot manipulation with in-the-wild human demonstrations. *arXiv preprint arXiv:2511.16661*, 2025.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

Yana Hasson, Gül 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 (CVPR)*, pp. 11807–11816, 2019.

Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *International Conference on 3D Vision (3DV)*, 2021.

Guanqi He and Wentao Zhang. WujiHand Retargeting. `https://github.com/wuji-technology/wuji-retargeting`, 2026.

Edmond S. L. Ho, Taku Komura, and Chiew-Lan Tai. Spatial relationship preserving character motion adaptation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 29(4), 2010.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

Taylor Howell, Nimrod Gileadi, Saran Tunyasuvunakool, Kevin Zakka, Tom Erez, and Yuval Tassa. Predictive sampling: Real-time behaviour synthesis with MuJoCo. *arXiv preprint arXiv:2212.00541*, 2022.

Jhen Hsieh, Kuan-Hsun Tu, Kuo-Han Hung, and Tsung-Wei Ke. DexMan: Learning bimanual dexterous manipulation from human and generated videos. *arXiv preprint arXiv:2510.08475*, 2025.

Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Lingyu Kong, Ruicheng Li, Ruicheng Wang, Sicheng Xu, Chengtang Yao, Jianfeng Xiang, and Jiaolong Yang. MoGe-3: Fine-detail monocular geometry estimation with self-guided sparse volumetric refinement. *arXiv preprint arXiv:2607.17967*, 2026.

Arjun S. Lakshmipathy, Jessica K. Hodgins, and Nancy S. Pollard. Kinematic motion retargeting for contact-rich anthropomorphic manipulations. *ACM Transactions on Graphics*, 2025.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Libin Liu, KangKang Yin, Michiel van de Panne, Tianjia Shao, and Weiwei Xu. Sampling-based contact-rich motion control. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 29(4), 2010.

Vincent Liu, Ademi Adeniji, Haotian Zhan, Siddhant Haldar, Raunaq Bhirangi, Pieter Abbeel, and Lerrel Pinto. EgoZero: Robot learning from smart glasses. *arXiv preprint arXiv:2505.20290*, 2025a.

Xueyi Liu, Kangbo Lyu, Jieqiong Zhang, Tao Du, and Li Yi. QuasiSim: Parameterized quasi-physical simulators for dexterous manipulations transfer. In *European Conference on Computer Vision (ECCV)*, 2024a.

Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. DexTrack: Towards generalizable neural tracking control for dexterous manipulation from human references. In *International Conference on Learning Representations (ICLR)*, 2025b.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025c.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-action-object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Tyler Ga Wei Lum, Olivia Y. Lee, C. Karen Liu, and Jeannette Bohg. Crossing the human-robot embodiment gap with sim-to-real RL using one human demonstration. In *Conference on Robot Learning (CoRL)*, 2025.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. DexMachina: Functional retargeting for bimanual dexterous manipulation. *arXiv preprint arXiv:2505.24853*, 2025.

Juncheng Mu, Sizhe Yang, Yiming Bao, Hojin Bae, Tianming Wei, Linning Xu, Boyi Li, Huazhe Xu, and Jiangmiao Pang. DexImit: Learning bimanual dexterous manipulation from monocular human videos. In *Robotics: Science and Systems (RSS)*, 2026.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3M: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022.

Bhawna Paliwal, Haritheja Etukuru, William Liang, Pieter Abbeel, Nur Muhammad Mahi Shafiullah, and Jitendra Malik. Do as I Do: Dexterous manipulation data from everyday human videos. *arXiv preprint arXiv:2606.19333*, 2026.

Chaoyi Pan, Changhao Wang, Haozhi Qi, Zixi Liu, Homanga Bharadhwaj, Akash Sharma, Tingfan Wu, Guanya Shi, Jitendra Malik, and Francois Hogan. SPIDER: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484*, 2025.

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

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

Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision (ECCV)*, pp. 570–587, 2022.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems (RSS)*, 2023.

Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3D deep learning with PyTorch3D. *arXiv preprint arXiv:2007.08501*, 2020.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 36(6), 2017.

Kenneth Shaw, Shikhar Bahl, and Deepak Pathak. VideoDex: Learning dexterity from internet videos. In *Conference on Robot Learning (CoRL)*, 2022.

Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. MimicPlay: Long-horizon imitation learning by watching human play. In *Conference on Robot Learning (CoRL)*, 2023.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and C. Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Robotics: Science and Systems (RSS)*, 2024.

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 (CVPR)*, pp. 5261–5271, 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.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 41(4), 2022.

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

Grady Williams, Paul Drews, Brian Goldfain, James M. Rehg, and Evangelos A. Theodorou. Aggressive driving with model predictive path integral control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2016.

Jane Wu, Georgios Pavlakos, Georgia Gkioxari, and Jitendra Malik. Reconstructing hand-held objects in 3D. *arXiv preprint arXiv:2404.06507*, 2024.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018.

Lujie Yang, Xiaoyu Huang, Zhen Wu, Angjoo Kanazawa, Pieter Abbeel, Carmelo Sferrazza, C. Karen Liu, Rocky Duan, and Guanya Shi. OmniRetarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3D reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3895–3905, 2022.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19717–19728, 2023.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Yuting Ye and C. Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 31(4), 2012.

Zhao-Heng Yin, Changhao Wang, Luis Pineda, Krishna Bodduluri, Tingfan Wu, Pieter Abbeel, and Mustafa Mukadam. Geometric retargeting: A principled, ultrafast neural hand retargeting algorithm. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OakInk2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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 (ECCV)*, 2020.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. HaWoR: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815, 2025.

A RECONSTRUCTION DETAILS

Table 6 lists the key settings of Section 3.1; ℓ is the index distal phalanx length. For the single-image benchmark of Section 4.1, the intrinsics come from the dataset and the MoGe-3 depth is used at its own metric scale, since a single frame allows neither frame alignment nor a hand-defined unit.

Table 6: Key reconstruction settings.

Component	Setting
Segmentation	SAM 3.1; objects are grounded by their category words and propagated through the clip
Depth and camera	MoGe-3 (ViT-G); focal length is the median estimate over three frames, principal point at the image center
Frame alignment	similarity to the first keyframe on table pixels within 400 px of the objects, 20% residual trimming
Object mesh	SAM 3D on the frame with at most 5% occlusion and the largest visible mask, decimated to 40k faces
Object pose	FoundationPose from the object mask in every frame, 5 refinement iterations
Hand placement	3 rounds of the similarity fit with 10% residual trimming; $\kappa = 2.41$, i.e. $w_i = 0.3$ at $d_i = \ell/2$
Refinement gate	only hands within 10 mm of the object are refined
Schedule	5 cycles of 10 Adam (Kingma & Ba, 2015) steps for each of the two stages; crossing repair after every cycle
Image alignment	learning rate 2×10^{-3} for $\boldsymbol{\theta}$ and 2×10^{-2} for $\boldsymbol{\beta}$; $\lambda_{\text{ord}} = 30$, $s_{\text{o}} = 5$ mm
Contact-penetration	learning rate 10^{-3} for $\boldsymbol{\theta}$ and $\mathbf{T}_h$; $\lambda_c = 66$, $\lambda_p = 7 \times 10^5$, $d_0 = 0.001\ell$, $\sigma_c = \sigma_p = 0.5\ell$
Crossing repair	2 mm clearance; mean vertex displacement at most 15 mm, 5 mm per cycle
Keyframes	sampling ratio $r = 0.2$

B ADAPTIVE KEYFRAME SAMPLING

A_t in Equation (1) climbs steeply during dynamic motions and stalls whenever the hand moves smoothly or holds still, crowding keyframes into reversals and transients, and spending almost none on steady stretches, maximizing action fidelity under a fixed compression ratio (Figure 3). Each hand accumulates its own A_t so a stationary hand does not pay for the other's keyframes.

Physics-in-the-loop refinement runs one CMA-ES search per keyframe, so the sampling ratio r sets its cost, measured in rollouts (Table 7). On XHand we compare adaptive sampling with uniform sampling of the same keyframe budget over 100 tasks at $r \in \{0.05, 0.1, 0.2, 1\}$, on the protocol of Section 4.2 (Table 7). At $r = 1$ the two rules coincide. At every lower ratio, adaptive sampling matches or exceeds uniform sampling on all three metrics with a comparable number of rollouts. Rollouts do not fall in proportion to r: each keyframe needs 30–50% more of them than at $r = 1$. With dense keyframes, the solution at one keyframe is nearly optimal for the next, so the warm-started search meets its stopping criterion quickly. As keyframes grow sparser, consecutive keyframes differ more, the previous solution becomes a poorer starting point, and each search needs more candidates to converge.

C ALGORITHMS

Algorithm 1 gives the finger-crossing repair of Section 3.1, and Algorithm 2 the contact-aware retargeting refinement of Section 3.2.

D SIMULATION SETTINGS

All methods are replayed in the same MuJoCo scene and scored on the same ground truth; only the hand trajectory differs. Table 8 lists the shared settings. Our method and OmniRetarget enter as keyframes interpolated with PCHIP (rotations with an $\mathrm{SO}(3)$ spline) on a 60 Hz grid, SPIDER as its commanded controls resampled from 100 Hz to 60 Hz (holding the last state if its rollout ends early),

Table 7: Keyframe sampling on XHand, 100 tasks, ATE criterion. At each ratio r both rules select the same number of keyframes and keep the first and last frame. Rollouts: number of physics simulations, i.e., candidate actions the optimizer tries out in simulation, which measures the actual compute (median over tasks).

	Adaptive				Uniform			
r	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Rollouts	Succ. (%) ↑	Trans. (cm) ↓	Rot. (°) ↓	Rollouts
0.05	**57**	**4.4**	**29**	0.9K	50	5.6	33	1.0K
0.1	**84**	**1.9**	**13**	1.9K	82	2.3	14	1.7K
0.2	**83**	**1.9**	**12**	3.4K	81	**1.9**	13	3.4K
1	–	–	–	–	**89**	**1.7**	**12**	**13K**

Algorithm 1 Finger-crossing repair of a MANO hand.

Require: MANO hand $(\boldsymbol{\theta}_0, \boldsymbol{\beta}, \mathbf{T}_{h,0})$, as placed or after a refinement cycle
n_f: number of points at which finger f's keypoint polyline γ_f (wrist–knuckle–middle–distal–tip) crosses the object surface; odd: the tip is inside, nonzero even: the finger passed through
F: fingers to repair
M: metric on pose changes; $\|\Delta\|_M^2$ is the squared pixel motion of all MANO vertices in the image when the pose changes by Δ

$F \leftarrow \{f : n_f > 0\}$
$(\boldsymbol{\theta}, \mathbf{T}_h) \leftarrow (\boldsymbol{\theta}_0, \mathbf{T}_{h,0})$ ▷ $\boldsymbol{\beta}$ fixed; joints of F and hand placement vary
while $F \neq \emptyset$ **do**
  $\rho_f \leftarrow L_f$ if $n_f > 0$, else d_f ▷ ρ_f: residual of finger f
    ▷ L_f: length of γ_f from its first entry point to the tip
    ▷ d_f: depth of f's deepest vertex inside the object
  $(\boldsymbol{\theta}^\star, \mathbf{T}_h^\star) \leftarrow$ the pose nearest to $(\boldsymbol{\theta}_0, \mathbf{T}_{h,0})$ under M at which every linearized ρ_f vanishes
  move $(\boldsymbol{\theta}, \mathbf{T}_h)$ toward $(\boldsymbol{\theta}^\star, \mathbf{T}_h^\star)$ by the largest $\alpha \leq 1$ that lowers $\sum_{f \in F} \rho_f$,
  $F \leftarrow \{f \in F : n_f > 0$ or f has vertices inside the object$\}$
return $(\boldsymbol{\theta}, \mathbf{T}_h)$, or $(\boldsymbol{\theta}_0, \mathbf{T}_{h,0})$ if the repair fails or moves the hand too far

and ManipTrans as its executed states, frame for frame. Every method's commands are projected onto the joint coupling of the underactuated hands before replay.

Table 8: Simulation settings shared by all methods in the transfer benchmark.

Component	Setting
Solver	timestep 1/240 s (4 steps per 60 Hz frame); implicit-fast integrator; Newton solver with 10 iterations and 10 no-slip iterations; elliptic friction cones, impedance ratio 10
Gravity	9.81 m/s^2 along the negative normal of the support plane
Contacts	condim 6; friction (1.0, 0.01, 0.0002); solref (0.003, 1), solimp (0.9, 0.99, 0.001); margin 1 mm; the hands do not collide with the table
Table	plane taken from the ground-truth scene
Objects	ground-truth initial pose and velocity; cataloged masses (typical weights for TACO and OakInk2, manufacturer weights for ARCTIC) spread uniformly over CoACD (Wei et al., 2022) hulls (2 mm, at most 32); objects that neither move nor are touched in the ground truth are fixed; every annotated object is present
Articulations	ARCTIC objects: limited hinge over the recorded range, damping 0.02 N m s, no friction loss
Hands	official kinematic models with their joint damping, friction loss and torque limits; position servos with $k_p = 500$, $k_v = 3$ and armature 0.001 for every hand; collision from CoACD hulls of the visual meshes (0.5 mm for Shadow, 2 mm otherwise); fingertip–fingertip and fingertip–palm self-collision
Wrist	free root welded to a mocap target, solref (0.0083, 1), solimp (0.99, 0.999, 0.001)
Control	targets interpolated with cubic splines (rotation spline for the root) to every step; the whole clip is replayed open-loop from its first frame

Algorithm 2 Contact-aware retargeting refinement.

Require: solver pose q^{init}; object samples O; vertices V_p, SDF ψ_p, anchors A_p and region R_p of p
M: metric on changes Δq of joints and wrist; $\|\Delta q\|_M^2$ is the weighted squared motion of the anchors, the wrist and the joints, with the joints weighted least
ρ_p: residual of p; $\rho = \left[\rho_{\text{thumb}}^\top, \rho_{\text{index}}^\top, \ldots, \rho_{\text{pinky}}^\top, \rho_{\text{palm}}^\top\right]^\top$; $J_\rho = \partial\rho/\partial q$
$S = (N_{\text{cross}}, \|\rho\|_1)$: the number of fingers with $n_p > 0$, then the summed residual, compared lexicographically ($<_{\text{lex}}$): fewer crossing fingers always ranks first

$q \leftarrow q^{\text{init}}$
repeat
  for $p \in \{\text{fingers, palm}\}$ **do**
    if $n_p > 0$ **then** ▷ p crosses the object: pull it out first (Algorithm 1)
      $\rho_p \leftarrow L_p$
    else ▷ p is outside: remove penetration and enclosure, reach the contact region
      $\rho_p \leftarrow \big(\text{ReLU}(-\phi(V_p)),\ \text{ReLU}(-\psi_p(O)),\ \text{distance}(A_p, R_p)\big)$
  solve $(J_\rho^\top J_\rho + \mu M)\,\Delta q = -J_\rho^\top \rho$ ▷ damping $\mu > 0$ keeps $\|\Delta q\|_M$ small
  $q \leftarrow q + \alpha\Delta q$ with the largest $\alpha \le 1$ that lowers S
until no violation remains or no step lowers S
return q if $S(q) <_{\text{lex}} S(q^{\text{init}})$, else q^{init}

E TRANSFER ABLATION DETAILS

Contact refinement's open-loop losses in Table 5 are mostly false successes: in the tasks where fingers pass fully through the object, the interlocked object follows the hand as it would in a real grasp and passes the ATE criterion without being held; the remaining losses come from contact lost between interpolated keyframes. After physics-in-the-loop refinement, contact refinement also lowers the fraction of frames with hand-object contact from 41.4% to 35.1%, while object errors fall on the Inspire Hand and SharpaWave and change by at most 0.1 cm and 1° on the other three hands. The contacts it removes are thus largely interlocking caused by penetration rather than contacts that grasp or support the object.

F END-TO-END FAILURE ANALYSIS

Both baselines estimate the camera with MoGe and replay the hand at its reconstructed position. TACO is filmed with a 24° telephoto lens, which MoGe reads as 42–52°, so the baselines place the whole scene at about half its true distance: although the hand pose itself is accurate, the wrist lands 0.73–0.93 m from its true position and the object moves by more than 1 cm in only 2 of 101 replays. Upstream image-based geometry estimation is thus the largest bottleneck of both baselines. OakInk2's 70° head-mounted camera lies within MoGe's prior and the wrist error falls to 37–59 mm, yet the hand pushes the object aside instead of grasping it, reproducing at most 3% of the ground-truth contacts, so grasping is a second bottleneck, independent of depth. **OmniHOI** fails on 28 clips, 14 in perception, 13 in execution and 1 in motion planning. The perception failures reconstruct the object at the wrong size from a mask that misses or overruns it (12 clips), at the wrong shape (1), or fail to ground it (1); in the execution failures the hand lets the object rotate (7), never closes on it (5), or drifts off the ground-truth path (1).

G REAL-WORLD DEPLOYMENT

Figures 9 and 10 show all seven real-world tasks of Section 4.4 as film strips. For each task, the top row is the input human video and the bottom row the execution on the two XHands; each row has seven frames evenly spaced in time over the motion in its video. The two videos of a task differ in length and are sampled independently.

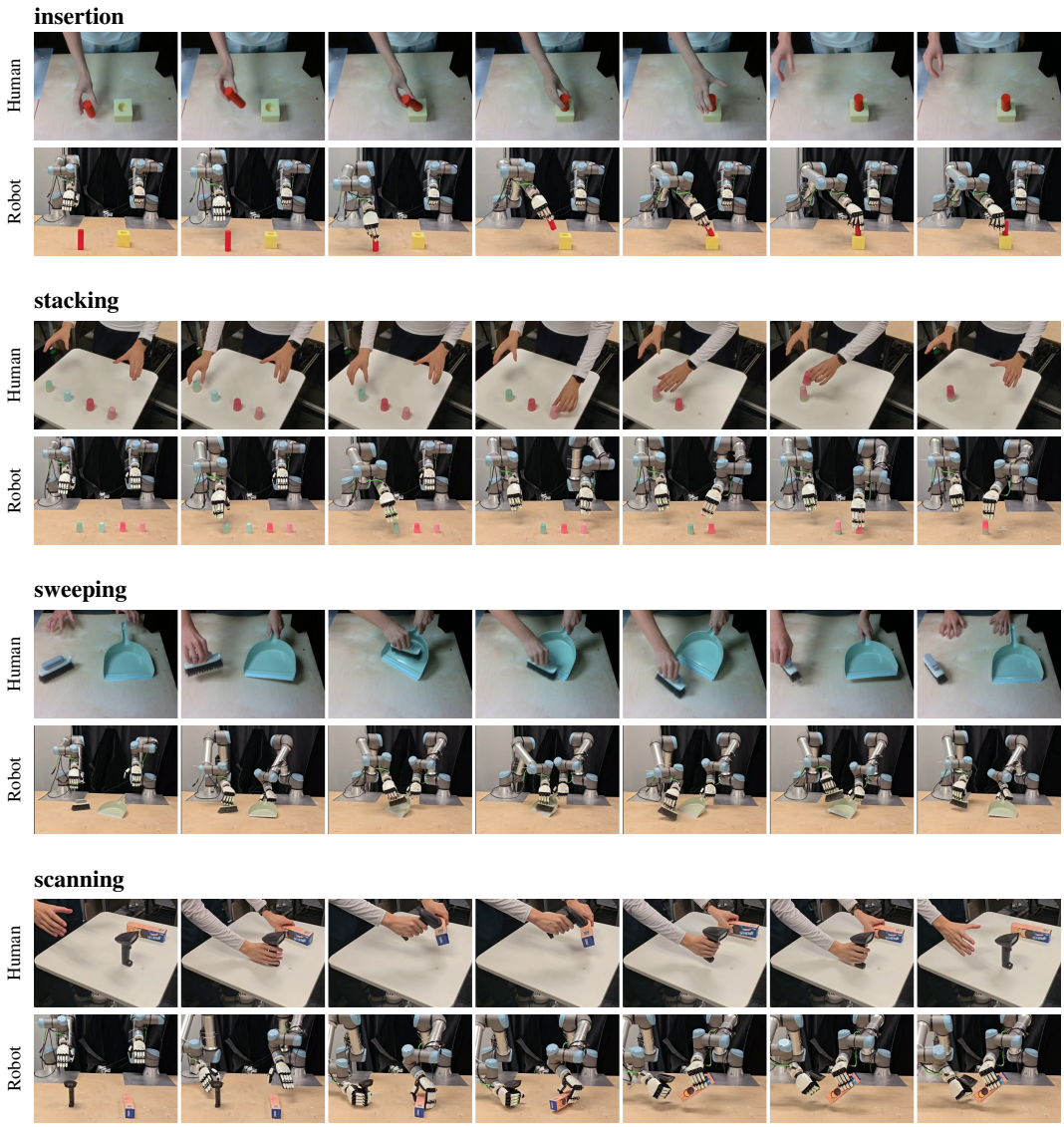


Figure 9: **Real-world film strips**: insertion, stacking, sweeping and scanning. For each task, top: the input human video; bottom: the robot executing the trajectory. Time runs from left to right.

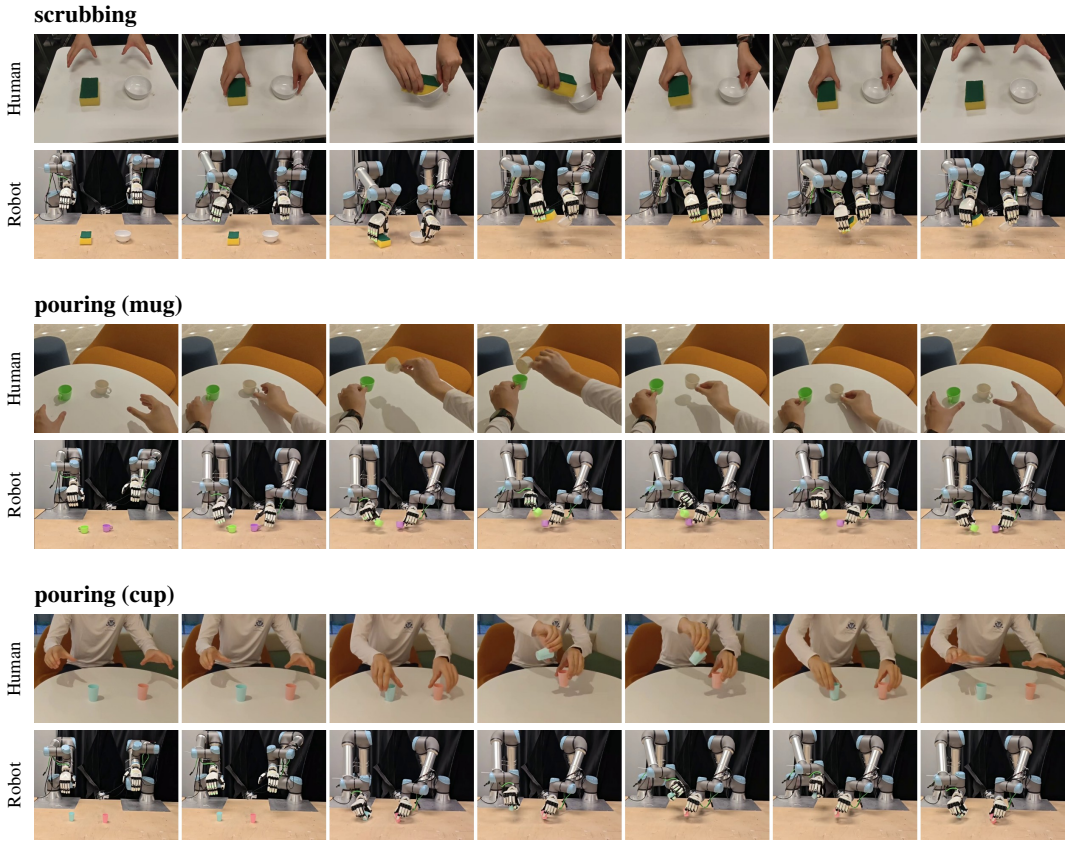


Figure 10: **Real-world film strips**, continued: scrubbing and the two pouring tasks, laid out as in Figure 9.