EGOMATRIX: PHYSICALLY FAITHFUL DEXTEROUS ROBOT DEMONSTRATIONS FROM EGOCENTRIC HUMAN VIDEOS

Anonymous authors
Paper under double-blind review

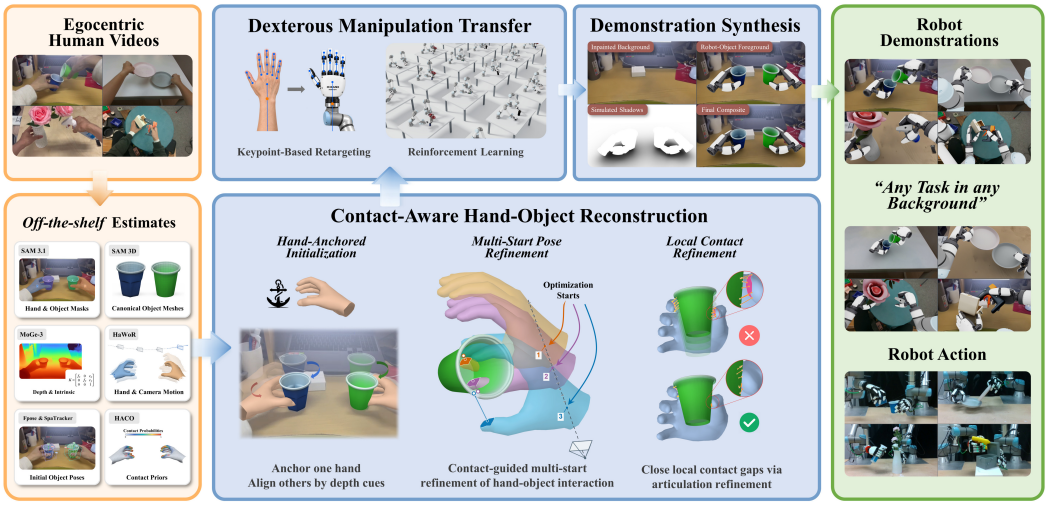


Figure 1: **Overview of EgoMatrix.** Starting from monocular egocentric human videos, EgoMatrix performs contact-aware hand-object reconstruction, dexterous manipulation transfer via retargeting and RL, and physically faithful robot observation synthesis. It produces robot trajectories executable on real hardware and synchronized observation videos, with the same trajectories transferable to diverse, task-compatible backgrounds.

ABSTRACT

Egocentric human videos offer a scalable alternative to robot teleoperation for collecting dexterous manipulation data. However, two challenges limit their use as reliable training data: accurate hand-object interaction (HOI) reconstruction from monocular RGB and high-quality robot demonstrations, especially visual observations. We present **EgoMatrix**, an end-to-end data engine that converts monocular egocentric human videos into bimanual dexterous trajectories and paired robot observation videos, with the trajectories directly replayable on real robots. It reconstructs HOI with emphasis on relative geometry and contact rather than exact metric scene recovery, then uses retargeting and reinforcement learning (RL) to reproduce the demonstrated tasks in simulation. We jointly render robot hands and objects from these simulated trajectories to produce physically faithful observations synchronized with state and control records. Rendering the same trajectories in diverse, task-compatible backgrounds expands paired training data while preserving robot-object motion and contact. Across TACO, OakInk2, and self-collected data, EgoMatrix achieves 54.84% end-to-end success in simulation versus 14.52% for the strongest baseline.Reconstruction reaches state-of-the-art performance on most metrics of the two public benchmarks. Qualitative results show more faithful hand-object interactions than prior methods, while feature-space analysis shows closer alignment with real-robot observations than GPT-6 Astra.

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.

1 INTRODUCTION

Learning generalizable dexterous manipulation policies requires large, diverse datasets pairing visual observations with robot actions (Kim et al., 2025). Teleoperation and motion capture provide high-quality demonstrations, but specialized hardware and operator effort make them costly to scale, especially for multi-fingered hands (Wang et al., 2024; Xu et al., 2025). Egocentric videos offer a more accessible source of diverse human manipulation (Paliwal et al., 2026), yet converting them into robot demonstrations of comparable quality remains challenging (Liu et al., 2026).

Recent systems have shown that manipulation can be transferred from monocular human videos to dexterous robots (Paliwal et al., 2026; Mu et al., 2026; Ren et al., 2026). However, different combinations of 3D scale, depth, and hand-object placement can produce similar 2D projections, making relative geometry and contact difficult to recover reliably from monocular images. A reconstruction may therefore match the input images while still containing hand-object gaps or interpenetration. Such errors can conflict with the contacts required for successful manipulation.

Generating visual observations consistent with the robot's actual manipulation poses a second challenge after successful task transfer in simulation. Data engines such as EgoEngine generate paired robot observations and trajectories (Liu et al., 2026), but retaining the original object imagery can leave depicted contacts and object motion inconsistent with the robot trajectory. For policy learning, visual observations should reflect how the robot actually manipulates the object.

We present **EgoMatrix**, an end-to-end data engine that converts monocular egocentric human videos into bimanual dexterous robot trajectories and synchronized observation videos. Our reconstruction prioritizes relative hand-object geometry and contact over exact metric scene recovery. We first use reconstructed hand geometry and relative-depth cues to establish the scene scale and initialize hand-object alignment. Multi-start pose refinement then corrects wrist and object misalignments, followed by local contact refinement to close contact gaps and reduce interpenetration. The refined trajectories guide retargeting and RL, enabling robots to complete the demonstrated tasks in simulation.

We then jointly render the robot hands and objects from successful simulated trajectories to produce physically faithful observations synchronized with state and control records. These observations are composited into the source scene with simulated shadows, while support-plane alignment and target-view rendering extend the same manipulation to diverse, task-compatible backgrounds.

This enables demonstrations from large egocentric video datasets to be converted into paired robot data for a specified real-world workstation, potentially reducing additional robot data collection when adapting pretrained policies to that workstation.

Our contributions are threefold:

- We present **EgoMatrix**, an end-to-end data engine that converts monocular egocentric human videos into bimanual dexterous robot demonstrations with synchronized observations and state-control trajectories, supporting direct trajectory replay on real robots. EgoMatrix achieves 54.84% end-to-end success in simulation versus 14.52% for the strongest baseline.
- We develop **contact-aware hand-object reconstruction** with multi-start pose refinement and local contact refinement, achieving state-of-the-art performance on most reconstruction metrics. Ablations show that both stages improve end-to-end success in simulation.
- We introduce **physically faithful robot observation synthesis** by jointly rendering robot hands and objects from successful simulated trajectories. Support-plane alignment and target-view rendering expand the paired data across diverse, task-compatible backgrounds while preserving robot-object motion and contact.

2 RELATED WORK

Monocular Hand-Object Reconstruction. Monocular hand-object reconstruction recovers 3D hand and object geometry from RGB videos, often by matching rendered reconstructions to image-space observations (Fan et al., 2024; Aboukhadra et al., 2026). Occlusion limits this evidence, motivating learned shape and generative priors for hidden object regions (Wang et al., 2025; Chen et al., 2026c). Since geometry completion alone does not resolve ambiguous hand-object placement

and contact, contact maps and temporally aggregated correspondences provide additional alignment cues (Grady et al., 2021; Ren et al., 2026). EgoMatrix similarly exploits contact cues, using contact-guided multi-start pose optimization under image-space constraints and subsequent local contact refinement to better preserve the HOI demonstrated in the human videos.

Human-to-Robot Manipulation Transfer. After reconstruction, human hand-object motion must be adapted to the robot embodiment while preserving the demonstrated interaction. Existing methods combine geometric retargeting (Qin et al., 2022; Zhang et al., 2026) with grasp and motion planning (Mu et al., 2026), reference-tracking reinforcement learning (Li et al., 2025; Ren et al., 2026), or sampling-based trajectory optimization (Paliwal et al., 2026) to obtain executable robot motion. EgoMatrix follows the reference-tracking approach, using contact-aware reconstructed trajectories as motion references for policy learning.

Robot Observation Synthesis. Robot observations can be synthesized by replacing human hands while retaining source object content (Zhang et al., 2026; Liu et al., 2026), or by generative models conditioned on images or simulated geometry (Chen et al., 2025; 2026a). These image-level approaches may leave depicted contacts or object motion inconsistent with the robot trajectory. A different line of work starts from recorded robot demonstrations and uses physics simulation to generate additional training data (Jiang et al., 2025; Fang et al., 2025). EgoMatrix instead starts from monocular human videos, transfers the demonstrated manipulation to successful robot trajectories in simulation, and jointly renders the robot hands and objects from those trajectories to produce physically faithful observations.

3 METHOD

Given a monocular egocentric human video, EgoMatrix reconstructs hand-object interaction, transfers the demonstrated task to a dexterous robot via retargeting and RL, and synthesizes synchronized robot observations and motion data (Fig. 1). The resulting trajectories can also be rendered in different task-compatible backgrounds.

3.1 CONTACT-AWARE HAND-OBJECT RECONSTRUCTION

Without any external calibration or scene-specific metric references, monocular RGB observations cannot uniquely determine metric scene geometry. We therefore target two properties: **manipulation-relevant 3D consistency** for motion transfer and **source-view consistency** for video compositing. Our reconstruction proceeds from hand-anchored initialization to multi-start pose refinement and local contact refinement.

Hand-Anchored Initialization We use SAM 3.1 (Carion et al., 2026) with natural-language prompts to obtain hand and object masks, and SAM 3D (Chen et al., 2026b) to reconstruct an unscaled canonical mesh $\mathcal{V}_j$ for each object. The temporal median of the per-frame intrinsics from MoGe-3 (Kong et al., 2026) defines a sequence-shared camera matrix $\mathbf{K}$, while HaWoR (Zhang et al., 2025) provides the MANO (Romero et al., 2017) hand trajectories.

We use the human-scale prior encoded by the reconstructed MANO geometry as the scene-scale anchor. Let $\mathcal{P}_{h,t}$ denote pixels shared by the visible hand mask and rendered MANO surface. We estimate a sequence-level scale from the MANO depth $z^H_{h,t}(u)$ and MoGe-3 depth $D^M_t(u)$:

$$\log s_{\text{seq}} = \mathcal{R}\left(\left\{\log z^H_{h,t}(u) - \log D^M_t(u)\right\}_{h,t,u}\right), \qquad \widetilde{D}_t = s_{\text{seq}} D^M_t, \tag{1}$$

where $\mathcal{R}$ robustly aggregates observations over $u \in \mathcal{P}_{h,t}$. Back-projecting the aligned depth $\widetilde{D}_t$ within each object mask to 3D yields an object point cloud, from which we estimate a sequence-shared uniform scale s_j for $\mathcal{V}_j$. We also use $\widetilde{D}_t$ as the metric-depth input to HaWoR's SLAM module, so that the camera translation is recovered under the same scale.

Using the scaled meshes, aligned depth, and camera trajectory, we initialize object poses with FoundationPose (Wen et al., 2024), augmented by rigid-motion hypotheses from SpatialTracker (Xiao et al., 2024) to improve propagation under hand occlusion.

To first reduce large relative-placement errors with a lightweight alignment before multi-start refinement, we fix the right hand and optimize the frame-wise translations of the left hand and both objects, while keeping hand articulation, object rotation, and object scale fixed:

$$\mathcal{L}_{\text{align}} = \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} + \underbrace{\lambda_{\text{ord}}\mathcal{L}_{\text{ord}} + \lambda_{\text{gap}}\mathcal{L}_{\text{gap}}}_{\mathcal{L}_{\text{depth}}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}}. \tag{2}$$

Here, $\mathcal{L}_{\text{2D}}$ encourages source-view consistency through visibility and silhouette constraints, while $\mathcal{L}_{\text{depth}}$ preserves pairwise front-back ordering and relative depth differences induced by $\widetilde{D}_t$; the remaining terms regularize the trajectory. We additionally estimate gravity and a shared support surface to ground the objects. Further details are given in Appendix A.1.

Multi-Start Pose Refinement After the hand-anchored initialization stage, we refine the wrist and object rigid poses while keeping MANO articulation fixed. HACO (Jung & Lee, 2025) provides contact probabilities $p_{h,t,i}$, from which we define likely contact vertices $\mathcal{I}_{h,t} = \{i \mid p_{h,t,i} > \tau_{\text{H}}\}$. For each MANO vertex $i \in \mathcal{I}_{h,t}$, we consider the camera ray passing through it. On concave or multi-layered objects, the ray may intersect multiple surfaces, and normal compatibility alone may still leave several plausible intersections on the object surface for the same hand vertex. Since pose refinement is sensitive to initialization (Grady et al., 2021), we construct multiple contact-guided starts that cover these alternative surface associations.

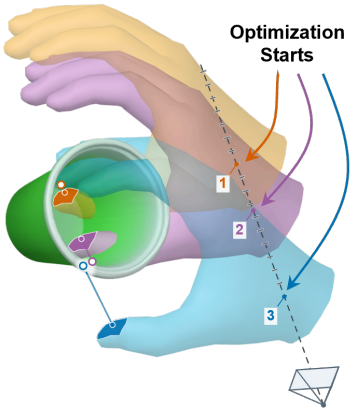


Figure 2: **Contact-guided multi-start initialization.** Different hand offsets associate candidate hand vertices with different points on the object surface, providing alternative starts for pose refinement.

As illustrated in Fig. 2, we generate starting poses by translating the hand along the ray from the camera center through the mean position of the initial MANO vertices. For each candidate k, we cast rays through the vertices in $\mathcal{I}_{h,t}$ and select the intersection closest to the corresponding vertex. These intersections are stored in object coordinates and define the targets of $\mathcal{L}_{\text{ray}}^{(k)}$, which attracts the hand vertices toward them during subsequent pose optimization.

To compare candidates, we group vertices in $\mathcal{I}_{h,t}$ into fixed patches using MANO mesh connectivity. For each patch, we record how often vertices are paired with the first, second, or later intersection along the ray. For each candidate, we compute its difference from every retained start and use the smallest difference as its diversity measure. Starting from the unshifted pose, we greedily select additional candidates with large diversity while penalizing large distances between paired points.

Each retained start jointly refines the wrist and bound-object poses. We collect the source-view, relative-depth, penetration, and regularization terms into $\mathcal{L}_{\text{base}}$. The objective is

$$\mathcal{L}_{\text{coarse}}^{(k,s)} = \mathcal{L}_{\text{base}} + \lambda_{\text{c}}\left[\eta_s\mathcal{L}_{\text{ray}}^{(k)} + (1-\eta_s)\mathcal{L}_{\text{surf}}\right], \tag{3}$$

where s indexes optimization steps and $\mathcal{L}_{\text{surf}}$ attracts vertices in $\mathcal{I}_{h,t}$ toward the current nearest object surface. The weight η_s decreases from 1 to 0, so all starts are ultimately refined under the same contact objective. Candidate selection and optimization details are provided in Appendix A.2.

Local Contact Refinement To improve the local hand-object geometry used for downstream manipulation transfer, we finally fix the object poses and jointly refine the wrist pose and MANO articulation. Temporally filtered HACO vertices are ray-cast to the object; for multiple intersections, we select those close to the hand vertex in depth with the strongest opposing normal match. These frame-wise contacts are stabilized in canonical object space following C2Dex (Ren et al., 2026). When stable patches provide insufficient opposing support, we supplement region-aware contacts according to the pad, palm, and side opposition types of the GRASP taxonomy (Feix et al., 2016).

For stable contact vertex i, let $\mathbf{v}_{t,i}$ and $\mathbf{y}^*_{t,i}$ denote the current hand vertex and its object-side target. For concise notation, we write the contact attraction as

$$\mathcal{L}_{\text{contact}} = \frac{1}{|\mathcal{T}_c|} \sum_{t \in \mathcal{T}_c} \sum_{i \in \mathcal{C}_t} w_{t,i} \left\| \mathbf{v}_{t,i} - \mathbf{y}^*_{t,i} \right\|_2^2, \tag{4}$$

where $\mathcal{C}_t$ is the stable contact set and $w_{t,i}$ is a contact weight derived from HACO confidence and temporal support. We combine $\mathcal{L}_{\text{contact}}$ with bidirectional penetration penalties and regularization terms to form the final contact refinement objective. Appendix A.3 provides the complete contact construction and optimization details.

3.2 Dexterous Manipulation Transfer

Given the reconstructed hand-object trajectories, we deliberately use a lightweight keypoint-based retargeting scheme rather than a more elaborate retargeting method. For each hand, we define fixed semantic correspondences between MANO joints and fingertips and robot-hand links. Let $\mathbf{x}_{t,\ell} \in \mathbb{R}^3$ denote the 3D position of MANO keypoint ℓ at frame t, and let $\mathbf{p}_t$, $\mathbf{R}_t$, and $\mathbf{q}_t$ denote the robot wrist position, wrist orientation, and finger configuration. With $\mathbf{f}_\ell(\mathbf{q}_t)$ denoting the wrist-local forward-kinematics position of the corresponding robot keypoint, we optimize

$$\mathcal{L}_{\text{ret}} = \frac{1}{TN} \sum_{t=1}^{T} \sum_{\ell=1}^{N} w_\ell \left\| \mathbf{p}_t + \mathbf{R}_t \mathbf{f}_\ell(\mathbf{q}_t) - \mathbf{x}_{t,\ell} \right\|_2 + \mathcal{R}_{\text{ret}}. \tag{5}$$

subject to the robot joint limits. $\mathcal{R}_{\text{ret}}$ regularizes wrist orientation, temporal smoothness, and table clearance. The two hands are retargeted independently, and the resulting wrist and joint trajectories are used to initialize physics-based policy learning.

We then adopt the RL framework of ManipTrans (Li et al., 2025), but omit its pretrained imitator and train a joint bimanual policy directly in Isaac Lab (Mittal et al., 2025). The retargeted robot trajectory is used for reference-state initialization, while the reconstructed MANO and object trajectories serve as tracking references. We train the policy with PPO (Schulman et al., 2017), using rewards for hand and object tracking, contact, and power. The learned policy is then rolled out in simulation to obtain a physically executable bimanual trajectory.

3.3 Demonstration Synthesis and Background Transfer

Following a successful policy rollout, we replay the recorded simulator states and jointly render the dexterous hands and reconstructed objects using the source-video camera intrinsics and frame-aligned camera poses. Because both the hands and objects are rendered from the same physics rollout, their relative geometry directly reflects the executed interaction. We remove the human hands, arms, and manipulated objects from the source video using SVOR (Hu et al., 2026) with removal masks derived from SAM 3.1 segmentation. We then estimate tabletop shadow attenuation from paired simulator renders with and without the robot-object foreground, and apply it to the cleaned background.

Let I_t^{sim} denote the simulator-rendered RGB image, M_t^{sim} the combined binary mask of the robot hands and manipulated objects, and $\widetilde{I}_t^{\text{bg}}$ the shadow-adjusted background. The synthesized frame is

$$I_t^{\text{syn}} = M_t^{\text{sim}} \odot I_t^{\text{sim}} + \left(1 - M_t^{\text{sim}}\right) \odot \widetilde{I}_t^{\text{bg}}. \tag{6}$$

The same successful execution can be transferred to another task-compatible background without modifying its internal hand-object geometry. For source and target task frames $\mathcal{F}_A$ and $\mathcal{F}_B$, respectively defined from the support surface and initial object placement of each scene, we compute a single rigid transform $\mathbf{T}_{B \leftarrow A} \in \mathrm{SE}(3)$ and apply it to the entire interaction. Thus, both hands and all manipulated objects preserve their relative poses, with no rescaling or frame-wise realignment. We re-render the same robot and object trajectories using the target camera trajectory and support-surface proxy. The rendered foreground is then composited with the cleaned target background using the same synthesis procedure. Different compatible backgrounds can therefore reuse the same

Table 1: End-to-end success rate of trajectory transfer across TACO, OakInk2, and self-collected demonstrations. Each entry reports the success rate and the number of successful trajectories over the total number of evaluated demonstrations.

Dataset	Tasks	Trajectory Transfer Success Rate (%) ↑		
		DexImit*	Do As I Do	EgoMatrix
TACO	brush, brush, bowl	7.14% (0.5/7)	28.57% (2/7)	**71.43% (5/7)**
	cut, knife, bowl	20.83% (2.5/12)	0.00% (0/12)	**33.33% (4/12)**
	screw, screwdriver, box	0.00% (0/11)	9.09% (1/11)	**45.45% (5/11)**
	smear, glue gun, box	0.00% (0/7)	0.00% (0/7)	**57.14% (4/7)**
OakInk2	stir spoon/rod	0.00% (0/8)	25.00% (2/8)	**62.50% (5/8)**
	rearrange tableware	0.00% (0/6)	16.67% (1/6)	**50.00% (3/6)**
Self-Collected	All 11 tasks	45.45% (5/11)	27.27% (3/11)	**72.73% (8/11)**
Overall	**All above tasks**	12.90% (8/62)	14.52% (9/62)	**54.84% (34/62)**

*For DexImit, a demonstration receives 0.5 credit when only one of the two hand-object trajectories satisfies the success criterion. Requiring both hands to succeed would give DexImit 0% on all public-dataset tasks.

robot and object trajectories and control records, changing only the common rigid transform applied to the trajectories, the camera viewpoint, and the visual background. Details of scene alignment and temporal synchronization are provided in Appendix B.2.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate on TACO (Liu et al., 2024), OakInk2 (Zhan et al., 2024), and a self-collected dataset. We use 4 TACO tasks with 37 demonstrations, 2 OakInk2 tasks with 14 demonstrations, and 11 self-collected tasks with one demonstration each. TACO and OakInk2 are used for reconstruction evaluation; all three datasets are used for end-to-end manipulation transfer.

Baselines. For end-to-end transfer, we compare with DexImit (Mu et al., 2026) and Do As I Do (Paliwal et al., 2026). For reconstruction, we compare with G-HOP (Ye et al., 2024), GHOST (Aboukhadra et al., 2026), MagicHOI (Wang et al., 2025), and ForeHOI (Chen et al., 2026c). For visual synthesis, we compare qualitatively with EgoEngine (Liu et al., 2026).

Metrics. For HOI reconstruction, we report root-aligned mean per-joint position error (MPJPE_{RA}) for hand pose; Chamfer distance (CD) and F-score at a 5 mm threshold (F@5) for object geometry; absolute trajectory error (ATE), absolute rotation error (ARE), relative translation error (RTE), and relative rotation error (RRE) for object trajectories (Shin et al., 2026); and hand-relative Chamfer distance (CD_h), simulation displacement (SD), and maximum penetration depth ($D_{\text{pen}}^{\max}$) for interaction quality (Fan et al., 2024; Jiang et al., 2021; Hasson et al., 2019). We also report the *Trajectory Transfer Success Rate*, defined as the fraction of demonstrations converted into robot trajectories that complete the same task in simulation. For TACO and OakInk2, we align the simulated and ground-truth object trajectories as in ATE and count a transfer as successful if the translation and rotation errors remain within 0.10 m and 1.0 rad throughout the task. For self-collected videos, where ground-truth object poses are unavailable, we manually judge each rollout and require both a valid grasp and completion of the intended manipulation.

4.2 End-to-End Video-to-Robot Trajectory Transfer

First, we evaluate end-to-end trajectory transfer on selected TACO and OakInk2 tasks and 11 self-collected demonstrations spanning diverse object geometries, grasp classes, and motion patterns. As shown in Table 1, EgoMatrix achieves the highest success rate across all evaluated task groups, reaching 54.84% (34/62) overall, compared with 14.52% (9/62) for Do As I Do and 12.90% (8/62) for DexImit. The gains are consistent across all three data sources and include tasks where the baselines rarely or never succeed, such as *screw, screwdriver, box* and *smear, glue gun, box*. Over-

Table 2: **Quantitative comparison of hand-object reconstruction on TACO and OakInk2.** Lower is better except for F@5. Best results are shown in bold and second-best results are underlined. "–" denotes metrics unavailable from the method outputs.

Dataset	Method	Object Geometry		Hand Pose	Object Trajectory				Interaction		
		CD ↓ (cm)	F@5 ↑ (%)	MPJPE_{RA} ↓ (mm)	ATE ↓ (cm)	ARE ↓ (°)	RTE ↓ (cm)	RRE ↓ (°)	CD_h ↓ (cm)	SD ↓ (cm)	$\text{D}_{\text{pen}}^{\max}$ ↓ (mm)
TACO	G-HOP	2.54	30.87	37.28	38.35	65.32	7.69	8.62	37.66	9.23	1.84
	MagicHOI	2.11	40.44	39.90	–	–	–	–	13.33	9.14	17.52
	ForeHOI	1.52	51.28	–	16.11	69.14	2.33	8.26	–	–	–
	GHOST	2.16	44.32	37.05	16.01	42.45	1.60	4.20	23.60	5.42	17.87
	DexImit	0.64	83.08	29.12	8.43	37.53	1.35	9.26	17.35	10.39	**0.35**
	Do As I Do	0.61	84.81	**22.16**	4.99	50.24	1.42	6.19	6.00	4.63	3.69
	EgoMatrix	**0.56**	**86.27**	23.07	**3.35**	**24.54**	**0.70**	**3.35**	**4.06**	**2.62**	2.64
OakInk2	G-HOP	1.55	42.36	27.39	46.08	63.19	3.44	6.79	**11.63**	9.51	0.45
	MagicHOI	1.95	37.59	23.52	–	–	–	–	16.85	**1.94**	15.71
	ForeHOI	1.00	58.36	–	9.31	83.09	1.43	7.13	–	–	–
	GHOST	1.61	50.78	29.13	8.09	**45.44**	1.16	4.00	18.63	2.54	21.18
	DexImit	1.20	71.47	25.27	12.55	78.75	1.57	10.66	30.71	9.66	**0.18**
	Do As I Do	0.59	84.61	24.04	8.93	70.19	1.79	5.19	16.52	4.91	1.70
	EgoMatrix	**0.57**	**87.75**	**20.47**	**2.52**	58.68	**0.56**	**2.70**	14.10	2.52	1.81



Figure 3: **Qualitative comparison of robot observation synthesis in the source background.** Each row compares Kiwi-Edit, GPT-6 Astra, EgoEngine, and EgoMatrix for the same source frame. Magnified regions highlight representative hand-object contacts. EgoMatrix produces more physically faithful robot observations with hand-object geometry consistent with the transferred manipulation.

all, EgoMatrix converts monocular human demonstrations into robot trajectories that reproduce the demonstrated manipulation with substantially higher success rates.

4.3 Reconstruction Quality

We next evaluate reconstruction quality, with particular attention to the hand-object interaction used for downstream transfer. As shown in Table 2, EgoMatrix achieves the best result on 8 of 10 metrics on TACO and 6 of 10 on OakInk2, including the best CD, F@5, ATE, and RTE on both datasets. It also achieves the best CD_h and SD on TACO and remains competitive on OakInk2.

Interaction metrics should be interpreted jointly. DexImit, for example, obtains the lowest $D_{\text{pen}}^{\max}$ on both datasets but substantially worse CD_h and SD, showing that low penetration alone does not imply accurate hand-object interaction. Fig. 4 further shows more interaction-consistent reconstructions than Do As I Do and DexImit on representative examples.

4.4 Visual Fidelity of Generated Robot Observations

Source-Background Fidelity. Fig. 3 compares robot observations synthesized in the original source background. Kiwi-Edit (Lin et al., 2026) largely preserves the scene and object placement but often fails to replace the human hands and arms with coherent robot geometry. Despite ex-

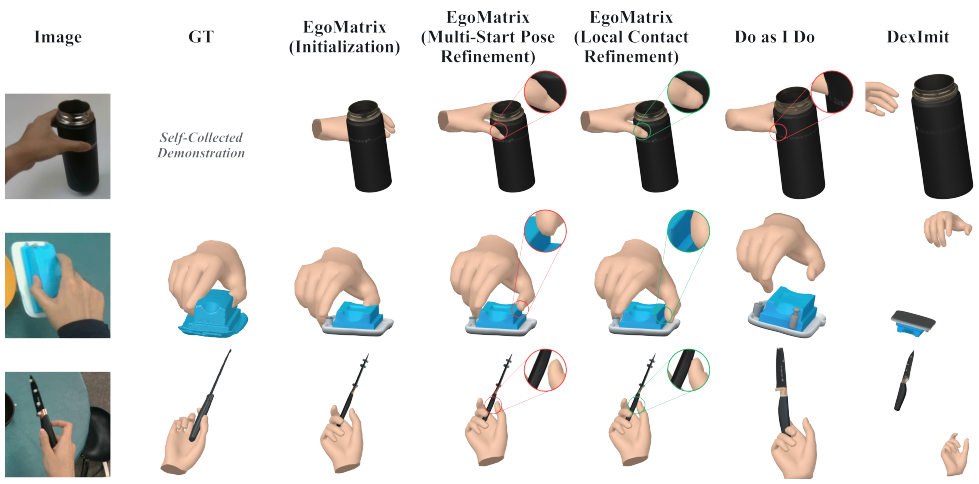


Figure 4: **Qualitative comparison of hand-object reconstruction.** Each row shows the same input frame across ground truth, the three stages of EgoMatrix, and competing methods. The three examples illustrate how Local Contact Refinement addresses three contact defects that can hinder downstream dexterous-hand transfer: **severe interpenetration** (top), **poorly localized contact** (middle), and **residual hand-object gaps** (bottom). Across these examples, EgoMatrix produces more interaction-consistent hand-object reconstructions than the competing methods.

Table 3: **Feature alignment (FD) with real robot observations.** FD over 52 frames per domain; lower is better.

Source	FD ↓			
	ResNet18	VGG16	DINOv2	DreamSim
Human	273.10	1003.64	468.06	1.0178
GPT6-Astra	153.83	804.60	**221.22**	0.6810
EgoMatrix	**100.23**	**414.66**	253.15	**0.4975**

Table 4: **Ablation of reconstruction refinement.** Each row cumulatively adds one stage, evaluating selected metrics.

Method	Self	TACO				
	Succ. ↑ (%)	Succ. ↑ (%)	MPJPE$_{\mathrm{RA}}$ ↓ (mm)	CD$_h$ ↓ (cm)	SD ↓ (cm)	D$_{\mathrm{pen}}^{\mathrm{max}}$ ↓ (mm)
Hand-Anchored Init.	18.18	5.41	**21.97**	7.14	4.00	3.82
+ Multi-Start Refinement	36.36	27.03	23.17	4.24	3.31	**1.51**
+ Local Contact Refinement	**72.73**	**48.65**	23.07	**4.06**	**2.62**	2.64

plicitly specifying a dual-UR5e system with XHand dexterous hands, GPT-6 Astra (OpenAI, 2026) still produces image distortion and mismatched robot hands or arms. EgoEngine better preserves robot appearance but exhibits local hand-object contact errors. EgoMatrix instead renders the robot hands and objects from the same successful simulation rollout, preserving the transferred interaction geometry while retaining the source background and viewpoint.

Target-Background Fidelity. We further compare GPT-6 Astra and EgoMatrix against ego-view real-robot observations in a target workspace. Using 52 frames from one manipulation sequence, we extract features with ResNet18 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2015), DINOv2 (Oquab et al., 2024), and DreamSim (Fu et al., 2023). Fig. 5 visualizes the feature distributions, while Table 3 reports Fréchet distance (FD) to the real-robot distribution in each full feature space; lower is better. EgoMatrix obtains the lowest FD for ResNet18, VGG16, and DreamSim, while GPT-6 Astra is slightly better on DINOv2. Qualitatively, Astra often preserves the source object's image-space size and placement under the viewpoint change and fails to reproduce the correct bimanual robot configuration. EgoMatrix re-renders the transferred robot and object trajectories from the target camera, explicitly adapting them to the new viewpoint. The fixed Astra prompt and generation protocol are provided in Appendix D.

4.5 Ablation Studies

Table 4 shows that both refinement stages are important for downstream manipulation transfer. The hand-anchored initialization already achieves an MPJPE$_{\mathrm{RA}}$ of 21.97 mm on TACO, outperforming all baselines in Table 2. The subsequent refinements slightly degrade this hand-pose metric but substantially improve interaction quality and end-to-end success: on TACO, success rises from 5.41% to 27.03% and then 48.65%, with the same trend on the self-collected tasks. This trade-off matches our goal of prioritizing manipulation-relevant hand-object geometry over hand pose accuracy alone.

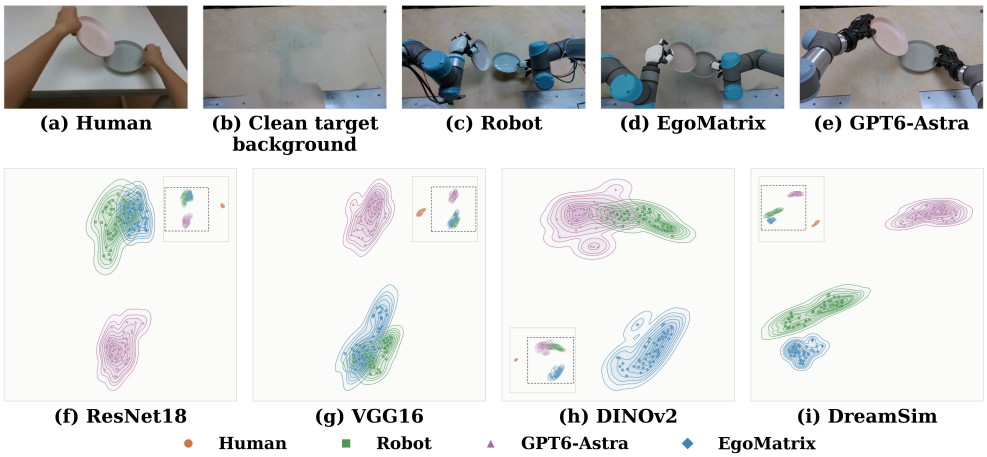


Figure 5: **Feature-space comparison for target-background transfer.** Top: source human frame, clean target background, real-robot observation, EgoMatrix, and GPT-6 Astra. Bottom: feature distributions from ResNet18, VGG16, DINOv2, and DreamSim over 52 frames per domain. EgoMatrix more closely matches the real-robot distribution in three of the four feature spaces.

Fig. 4 further illustrates the two stages. Multi-Start Pose Refinement corrects large initial pose errors in the first two examples while preserving the plausible initialization in the third. Local Contact Refinement then addresses interpenetration, poorly localized contact, and residual hand-object gaps. Together with the improved CD_h, SD, and transfer success, these results show the importance of interaction geometry for downstream transfer.

4.6 Real-World Trajectory Replay

We further validate the transferred trajectories on two UR5e arms equipped with XHand dexterous hands. From the trajectories generated by EgoMatrix, we select representative tasks for hardware replay, with source demonstrations drawn from both self-collected videos and public egocentric datasets. As shown in Fig. 6, the transferred trajectories complete the corresponding manipulations on the physical system. Additional hardware details and full trajectory sequences are provided in Appendix E. More video demonstrations are available on our anonymous project page: `https://egomatrix2027.github.io/`.



Figure 6: **Real-world trajectory replay.** Each adjacent pair shows a source human demonstration frame and the corresponding real-robot replay for the same task.

5 Conclusion and Limitations

We presented **EgoMatrix**, an end-to-end data engine that converts monocular egocentric human videos into bimanual dexterous robot demonstrations. Contact-aware reconstruction improves the hand-object references used for manipulation transfer, while joint rendering of simulated robot and object trajectories produces physically faithful visual observations and supports data generation across task-compatible backgrounds. Experiments demonstrate improved end-to-end manipulation transfer, strong reconstruction performance, and direct trajectory replay on real robots.

Limitations. EgoMatrix remains sensitive to errors from monocular perception and off-the-shelf estimators. Background transfer currently assumes a task-compatible scene with a configured support surface and does not re-simulate the manipulation against arbitrary target-scene geometry. Finally, while EgoMatrix generates synchronized visual and motion data, we have not yet evaluated their benefit for downstream policy or VLA training.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and language of the manuscript, to assist with literature and technical-resource retrieval, and to support research ideation and execution through brainstorming, methodological discussion, implementation and debugging assistance, and experimental planning. All AI-assisted suggestions, references, code, and research decisions were reviewed and verified by the authors. The authors take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses publicly available human-manipulation datasets and self-collected videos for research on dexterous robot learning. Our method uses only task-relevant hand-object motion and does not infer or use sensitive personal attributes. Public datasets are used according to their intended research purposes. The generated robot demonstrations are intended for research use; real-robot experiments are conducted in controlled settings, and deployment around people requires appropriate safety measures.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our results throughout the main paper and appendix. Section 3 describes the EgoMatrix pipeline and optimization objectives, while the appendix provides implementation details and training and simulation settings. Section 4.1 specifies the datasets, baselines, and evaluation metrics, and Appendix C gives the detailed definitions, alignment rules, aggregation procedures, and validity criteria used for reconstruction evaluation. Public datasets and baseline implementations are cited in the corresponding sections, and additional task-level details and qualitative results are included in the appendix and supplementary material.

REFERENCES

Ahmed Tawfik Aboukhadra, Marcel Rogge, Nadia Robertini, Abdalla Arafa, Jameel Malik, Ahmed Elhayek, and Didier Stricker. Ghost: Fast category-agnostic hand-object interaction reconstruction from rgb videos using gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 3394–3404, June 2026.

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 *International Conference on Learning Representations (ICLR)*, 2026.

Jason Chen, I-Chun Arthur Liu, Gaurav Sukhatme, and Daniel Seita. CRAFT: Video diffusion for bimanual robot data generation. In *International Conference on Intelligent Robots and Systems (IROS)*, 2026a.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. RoVi-Aug: Robot and viewpoint augmentation for cross-embodiment robot learning. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 209–233. PMLR, 2025. URL https://proceedings.mlr.press/v270/chen25a.html.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J. Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7220–7232, 2026b.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. Forehoi: Feed-forward 3d object reconstruction from daily hand-object interaction videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8868–8879, June 2026c.

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, June 2024.

Yu Fang, Yue Yang, Xinghao Zhu, Kaiyuan Zheng, Gedas Bertasius, Daniel Szafir, and Mingyu Ding. ReBot: Scaling robot learning with real-to-sim-to-real robotic video synthesis. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11351–11358, 2025. doi: 10.1109/IROS60139.2025.11246305.

Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M. Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on Human-Machine Systems*, 46 (1):66–77, 2016. doi: 10.1109/THMS.2015.2470657.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, pp. 50742–50768, 2023.

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, June 2021.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11807–11816, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, June 2016. doi: 10.1109/CVPR.2016.90.

Jiagao Hu, Yuxuan Chen, Fuhao Li, Zepeng Wang, Fei Wang, Daiguo Zhou, and Jian Luan. From ideal to real: Stable video object removal under imperfect conditions. *arXiv preprint arXiv:2603.09283*, 2026.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11107–11116, 2021.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. DexMimicGen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11639–11646, 2025. doi: 10.1109/ICRA55743.2025.11127809.

Daniel Sungho Jung and Kyoung Mu Lee. Learning dense hand contact estimation from imbalanced data. *Advances in Neural Information Processing Systems*, 38:133424–133457, 2025. doi: 10.52202/085713-4017.

Daniel Sungho Jung and Kyoung Mu Lee. Training-free dense hand contact estimation with multimodal large language models. *arXiv preprint arXiv:2605.05886*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In Pulkit Agrawal, Oliver Kroemer, and Wolfram Burgard (eds.), *Proceedings of The 8th Conference on Robot Learning*, volume 270

of *Proceedings of Machine Learning Research*, pp. 2679–2713. PMLR, 2025. URL https://proceedings.mlr.press/v270/kim25c.html.

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 e-prints*, pp. arXiv–2607, 2026.

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)*, pp. 6991–7003, 2025.

Yiqi Lin, Guoqiang Liang, Ziyun Zeng, Zechen Bai, Yanzhe Chen, and Mike Zheng Shou. Kiwi-edit: Versatile video editing via instruction and reference guidance, 2026.

Yangcen Liu, Shuo Cheng, Xinchen Yin, Woo Chul Shin, Alfred Cueva, Yiran Yang, Zhenyang Chen, Chuye Zhang, and Danfei Xu. Egoengine: From egocentric human videos to high-fidelity dexterous robot demonstrations. *arXiv preprint arXiv:2606.12604*, 2026.

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)*, pp. 21740–21751, June 2024.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, et al. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 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 *Proceedings of Robotics: Science and Systems*, Sydney, Australia, July 2026. doi: 10.15607/RSS.2026.XXII.003.

OpenAI. GPT-6 Astra System Card, September 2026. OpenAI Deployment Safety Hub.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

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.

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 *Computer Vision – ECCV 2022*, volume 13699 of *Lecture Notes in Computer Science*, pp. 570–587. Springer, 2022. doi: 10.1007/978-3-031-19842-7_33.

Jie Ren, Zhehao Jiang, Yinhong Yang, Haorui Jia, Han Jiang, Ben Li, Yao Yao, Cheng Lin, Qiu Shen, Zhenshan Bing, Xiao-Xiao Long, and Xun Cao. C2Dex: Contact-consistent reconstruction and retargeting for dexterous manipulation from monocular video. *arXiv preprint arXiv:2608.07045*, 2026.

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): 245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Jisu Shin, Junoh Lee, JunGyu Lee, Inhwan Bae, Dohyeon Lee, Hokyun Im, Youngwoon Lee, and Hae-Gon Jeon. ComPose: When to trust hands for object pose tracking. *arXiv preprint arXiv:2605.23523*, 2026.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015.

Chen Wang, Haochen Shi, Weizhuo Wang, Ruohan Zhang, Li Fei-Fei, and Karen Liu. DexCap: Scalable and portable mocap data collection system for dexterous manipulation. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.043.

Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. MagicHOI: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5957–5968, October 2025.

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

Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker: Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20406–20417, 2024.

Mengda Xu, Han Zhang, Yifan Hou, Zhenjia Xu, Linxi Fan, Manuela Veloso, and Shuran Song. DexUMI: Using human hand as the universal manipulation interface for dexterous manipulation. In Joseph Lim, Shuran Song, and Hae-Won Park (eds.), *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 437–459. PMLR, 2025. URL `https://proceedings.mlr.press/v305/xu25b.html`.

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)*, pp. 1911–1920, 2024. doi: 10.1109/CVPR52733.2024.00187.

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)*, pp. 445–456, June 2024. doi: 10.1109/CVPR52733.2024.00050.

Gu Zhang, Qicheng Xu, Haozhe Zhang, Jianhan Ma, Long He, Yiming Bao, Zeyu Ping, Zhecheng Yuan, Chenhao Lu, Chengbo Yuan, Tianhai Liang, Xiaoyu Tian, Maanping Shao, Feihong Zhang, Mingyu Ding, Yang Gao, Hao Zhao, Hang Zhao, and Huazhe Xu. Unidex: A robot foundation suite for universal dexterous hand control from egocentric human videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1841–1852, June 2026.

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.

APPENDIX

A ADDITIONAL DETAILS OF CONTACT-AWARE HAND-OBJECT RECONSTRUCTION

A.1 HAND-ANCHORED INITIALIZATION

Scale registration and camera consistency. For hand h at frame t, let $\mathcal{P}_{h,t}$ contain pixels supported by both the observed visible-hand mask and the rendered MANO surface. The sequence-level

scale in Eq. 1 is estimated by robustly aggregating

$$r_{h,t}(u) = \log z^{H}_{h,t}(u) - \log D^{M}_{t}(u), \qquad u \in \mathcal{P}_{h,t}, \tag{7}$$

over valid observations from both hands and all frames. The resulting factor s_{seq} is applied to the MoGe-3 depth and point map.

For each object, we back-project the aligned depth within its visible masks to 3D and robustly estimate its observed spatial extent over multiple frames. A single uniform scale s_j is then applied to the canonical mesh $\mathcal{V}_j$ for the entire sequence. The same aligned depth is also provided as the metric-depth input to the DROID-SLAM module used by HaWoR, so that camera translation is recovered under the same scale as the reconstructed hands and objects. Object tracking is subsequently performed using the scaled meshes and aligned scene geometry.

Object pose initialization and tracking. FoundationPose (Wen et al., 2024) initializes each object in its first valid frame and refines subsequent pose hypotheses. For later frames, we combine a camera-motion propagation hypothesis with rigid-motion hypotheses fitted from SpatialTracker (Xiao et al., 2024) point tracks. FoundationPose refines and scores these hypotheses to select the frame-wise object pose. Under severe hand occlusion, we avoid unconstrained full-pose recovery: translation-only refinement is used when sufficient image evidence remains, while otherwise the propagated pose is retained. This reduces abrupt pose changes when most of the object surface is unobserved.

Source-view alignment. The lightweight alignment in Eq. 2 jointly considers the two hands and two manipulated objects. For $\mathcal{L}_{\text{2D}}$, each entity is rendered independently and combined using depth-aware soft visibility. The loss compares the resulting visible regions with the observed masks using semantic, silhouette, boundary, and per-entity coverage terms. Ambiguous mask overlaps and invalid hand observations are excluded. When an object projection loses most of the support from its observed mask, an additional projection-capture term discourages the optimizer from moving the object out of view.

Pairwise relative-depth constraints. We partition reliable visible regions into local patches and compare the relative depths of patch pairs. Let $Z_{a,p}$ and $Z_{b,q}$ denote their current rendered depths and let $\bar{D}_{a,p}$ and $\bar{D}_{b,q}$ denote robust estimates from the aligned MoGe-3 depth. We define

$$\Delta Z_{ap,bq} = Z_{b,q} - Z_{a,p}, \qquad \Delta D_{ap,bq} = \bar{D}_{b,q} - \bar{D}_{a,p}. \tag{8}$$

The ordering term penalizes disagreement with the observed front–back relation,

$$\ell^{ap,bq}_{\text{ord}} = w^{ap,bq}_{\text{ord}} \rho\left(\frac{[m_{\text{ord}} - \text{sign}(\Delta D_{ap,bq})\Delta Z_{ap,bq}]_{+}}{\sigma_{\text{ord}}}\right), \tag{9}$$

while the gap term penalizes disagreement in relative depth,

$$\ell^{ap,bq}_{\text{gap}} = w^{ap,bq}_{\text{gap}} \rho\left(\frac{\Delta Z_{ap,bq} - \Delta D_{ap,bq}}{\sigma_{\text{gap}}}\right). \tag{10}$$

The two losses are aggregated over valid hand-hand, hand-object, and object-object patch pairs. Missing rendered support receives a fixed penalty rather than being ignored, preventing the optimizer from reducing the loss by moving an entity outside the observed region.

Temporal regularization and support grounding. Frame-wise translation corrections are regularized temporally after conversion to the common world frame, while a weak prior discourages unnecessary deviation from the initialization. We estimate the world-up direction from dominant background surface normals and infer a shared support height from intervals in which the objects are not in hand contact. Stable support configurations are selected over continuous intervals, after which object poses are adjusted to the inferred support surface. The hand trajectories are unchanged during grounding.

A.2 MULTI-START POSE REFINEMENT

Starting-pose selection. We generate candidates independently for each hand-object pair. All starts share an object baseline obtained by a short pose refinement under image, relative-depth, and support constraints. We use a HACO probability threshold of 0.8 and scan hand translations over $[-85, 85]$ mm at 5 mm intervals, retaining five starts per frame with contact evidence, including the unshifted pose.

The MANO patches remain fixed throughout each scan. For candidate k and patch p, we compute a histogram $\mathbf{h}_{k,p}$ of the selected intersection ranks along the camera rays, normalized by the number of vertices with valid intersections. We also record the fraction $f_{k,p}$ of rays that hit the object. The histogram uses nine bins: ranks $0, \dots, 7$ and a final bin for ranks ≥ 8; patches without hits have zero histograms. These quantities form the comparison vector $\mathbf{s}_k = \mathrm{concat}_p[\mathbf{h}_{k,p}, f_{k,p}/4]$.

Let $\mathcal{S}$ contain the indices of retained starts, initially only the zero-offset candidate. While candidates remain sufficiently different, we select

$$k^\star = \arg\max_{k \notin \mathcal{S}} \left[\min_{\ell \in \mathcal{S}} \frac{\|\mathbf{s}_k - \mathbf{s}_\ell\|_1}{\dim(\mathbf{s}_k)} - \lambda_q \widetilde{q}_k \right], \qquad \mathcal{S} \leftarrow \mathcal{S} \cup \{k^\star\}, \tag{11}$$

where $\lambda_q = 0.05$. The quality cost is $q_k = f_k \bar{g}_k + (1 - f_k) c_{\mathrm{miss}}$, where f_k is the overall ray-hit fraction, $\bar{g}_k$ is the mean distance between valid vertex-intersection pairs, and c_{miss} is a fixed cost for missed rays. We obtain $\widetilde{q}_k$ by min–max normalization over the entire scan, setting it to zero when all costs are equal. When the remaining candidates have negligible diversity, we instead select by increasing q_k, preferring offsets at least 10 mm from all retained offsets. If none satisfies this spacing, all remaining offsets are considered.

Candidate optimization. Each candidate jointly optimizes the wrist and its bound object, while the two hand-object pairs are solved independently. Translation corrections are defined in the camera frame, and rotation corrections left-multiply the initial orientation. MANO articulation and shape, object geometry and scale, and camera parameters remain fixed. Translation and rotation corrections are bounded by 100 mm and 10° for the wrist, and 50 mm and 12° for the object.

The source-view term in $\mathcal{L}_{\mathrm{base}}$ combines silhouette overlap and boundary alignment, using the other entity masks to exclude occluded regions. Relative-depth constraints follow Appendix A.1, but compare only the current hand-object pair using $\widetilde{D}_t$. Bidirectional penetration penalties limit interpenetration; table constraints preserve object support and discourage hand-table penetration.

We use the contact transition in Eq. 3. Ray targets remain fixed in object coordinates and follow the current object pose; vertices without valid ray targets use nearest-surface distance. Normal compatibility is introduced as the surface contact term replaces the ray term. Temporary pose anchors are reduced with the ray term. Temporal regularization is disabled during independent candidate optimization.

Candidate validation and temporal selection. After independent optimization, we first discard numerically invalid candidates and check contact and penetration quality. Candidates satisfying the contact-distance criterion are preferred, while penetration checks prevent strongly colliding solutions from dominating the selection. If no fully safe candidate is available, we retain the finite candidate with the smallest violation rather than discarding the frame. The zero-offset start receives no special preference, and candidates are not required to improve over the original initialization under a single global objective.

The remaining candidates are evaluated using a common geometric criterion that includes source-view consistency, relative depth, contact geometry, penetration, and support constraints. Candidate-specific ray targets and temporary initialization anchors are excluded from this score. Let U_t^k denote the resulting unary cost for candidate k at frame t. We then select a temporally coherent path using second-order dynamic programming:

$$\{k_t^\star\} = \arg\min_{\{k_t\}} \left[\sum_t U_t^{k_t} + \sum_t E_{\mathrm{pair}}(k_t, k_{t+1}) + \sum_t E_{\mathrm{acc}}(k_{t-1}, k_t, k_{t+1}) \right]. \tag{12}$$

Here, E_{pair} penalizes inconsistent relative hand-object states across adjacent frames, while E_{acc} penalizes abrupt wrist and object motion. These terms depend on the optimized geometry rather than

on changes in candidate indices. The selected path is finally refined with temporal regularization in the common world frame.

A.3 Local Contact Refinement

Stable contact construction. We follow the canonical-space stabilization strategy of C2Dex (Ren et al., 2026), while using temporally filtered HACO predictions to select the initial hand vertices. Small disconnected components are removed before ray casting. For each candidate vertex, we cast a ray from the camera center through its initial 3D position. Among the intersections with the bound object, we retain those within 15 mm in ray depth and with opposing-normal compatibility $-\mathbf{n}_h^\top \mathbf{n}_o > 0.5$, then select the intersection with the strongest normal match.

The selected points are transformed to canonical object coordinates and grouped into locally stable interaction segments according to hand articulation and relative hand-object motion. For each MANO vertex within a segment, we cluster its canonical observations using DBSCAN with $\epsilon = 5$ mm and $\mathrm{min_samples} = 3$. The dominant non-noise cluster is retained, and its medoid defines the stable object-side target. Using the medoid keeps the target on an observed object surface rather than introducing an averaged point away from the mesh.

Opposition supplementation. When the stable contacts do not provide sufficient opposing support, we supplement them according to the pad, palm, and side opposition types of the GRASP taxonomy (Feix et al., 2016). We use a simplified semantic partition of MANO adapted from the hand-part regions of ContactPrompt (Jung & Lee, 2026). Depending on the opposition type, supplementary vertices are drawn from the corresponding thumb, finger-pad, palm, or side regions and prioritized by HACO confidence. Their object-side targets are obtained from the nearest compatible surface point subject to distance and normal constraints. Supplementary contacts are introduced only when the existing contacts provide insufficient opposition; touch interactions do not use this mechanism.

Optimization. The object trajectory and MANO shape remain fixed. For each hand, we optimize the wrist translation, wrist rotation, and all 15 MANO joint rotations over the sequence, initialized from the Multi-Start Pose Refinement result. In implementation, the contact term in Eq. 4 is aggregated over contact patches:

$$\mathcal{L}_{\text{contact}} = \frac{1}{|\mathcal{T}_c|} \sum_{t \in \mathcal{T}_c} \sum_{p} W_{t,p} \frac{\sum_{i \in \mathcal{P}_{t,p}} a_{t,i} \left\| \mathbf{v}_{t,i} - \mathbf{y}^*_{t,i} \right\|_2^2}{\sum_{i \in \mathcal{P}_{t,p}} a_{t,i}}, \tag{13}$$

where $\mathcal{P}_{t,p}$ is a contact patch. The vertex weight $a_{t,i}$ combines HACO confidence and temporal support, while $W_{t,p}$ controls the contribution of each patch. This prevents densely sampled contact regions from dominating the refinement.

We penalize penetration in both directions. For hand-to-object penetration, let $d_{o,t}(\mathbf{v})$ denote signed distance to the object surface, positive outside:

$$\mathcal{L}_{h \to o} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{1}{N_h} \sum_{i} \left[-d_{o,t}(\mathbf{v}_{t,i}) \right]_+ . \tag{14}$$

The reverse term penalizes sampled object-surface points that lie inside the closed hand mesh. Additional regularization limits deviation from the coarse wrist and articulation parameters, penalizes second-order temporal variation, enforces anatomical joint-angle plausibility, and prevents hand-table penetration.

We optimize each hand independently with Adam for 1000 iterations. Wrist translation, wrist rotation, and articulation use learning rates 0.001, 0.005, and 0.0055, with residual bounds of 60 mm, 0.35 rad, and 0.45 rad per MANO joint, respectively.

B Details of Demonstration Synthesis and Background Transfer

B.1 Demonstration Synthesis

Background removal and shadow transfer. The removal masks combine the reconstructed hand and object masks with additional arm and sleeve segmentations. We apply video inpainting with SVOR (Hu et al., 2026) and use its native prediction as the background. The simulation foreground is the union of the two dexterous-hand and manipulated-object segmentation masks.

To preserve interaction shadows, we render the scene with and without the simulated hands and objects under identical camera and lighting conditions. Their linear-image intensity ratio defines a table-region shadow transmittance τ_t. We apply the resulting attenuation to the inpainted background before the binary compositing in Eq. 6. The foreground itself is not geometrically warped or modified by a generative model.

Temporal correspondence. The synthesized video is rendered at 30 Hz, whereas policy control and physics simulation run at 60 Hz and 120 Hz, respectively. Each video frame stores its associated simulator-state and source-camera indices, while the complete control-rate trajectory is retained separately. The recorded trajectory contains robot and object states, policy actions, actuator targets, wrist wrenches, and transition outcomes.

B.2 Background Transfer

Scene placement. Let A and B denote the source and target task coordinate systems. For each scene, we define a table-aligned task frame $\mathbf{F}$ whose origin is the midpoint of the two initial object positions projected onto the support plane. Its x-axis follows the direction between the two objects, while the z-axis is the world-up direction. The initial source-to-target placement is

$$\mathbf{T}^0_{B\leftarrow A} = \mathbf{F}_B \mathbf{F}_A^{-1}. \tag{15}$$

An optional fixed in-plane translation $\mathbf{D}_A$ may be applied for framing,

$$\mathbf{T}_{B\leftarrow A} = \mathbf{T}^0_{B\leftarrow A} \mathbf{D}_A. \tag{16}$$

The same transform is shared by both hands and all manipulated objects. No scale change or frame-wise alignment is applied.

Rather than rewriting the recorded source states, we equivalently express the target camera in the source task frame. If $\mathbf{C}_B(j)$ is the target camera-to-world pose at target frame j, the rendering camera for output frame i is

$$\mathbf{C}_A(i) = \mathbf{T}^{-1}_{B\leftarrow A} \mathbf{C}_B(j(i)) \,. \tag{17}$$

Target intrinsics are used for rendering, and the target support proxy is transformed consistently with the camera and scene placement.

Temporal mapping. The source execution remains on its original 60 Hz state and control trajectory, while the synthesized video is rendered at 30 Hz. We therefore use every second recorded source state. When the available target-background segment is shorter than the source output, target RGB frames and camera poses are sampled uniformly over the available segment using the same frame mapping; the source execution itself is not time-warped.

The target support representation is taken from the configured background scene rather than re-estimated during background transfer. Accordingly, this stage preserves the original executed hand-object geometry and its motion relative to a compatible support plane, but does not re-simulate contact dynamics in the target scene.

C Reconstruction Evaluation Protocol

We report ten reconstruction metrics: MPJPE_{RA}, CD, F@5, ATE, ARE, RTE, RRE, CD_h, SD, and $D^{\max}_{\text{pen}}$. Unless otherwise specified below, all methods are evaluated using the same metric implementation and aggregation protocol.

Hand pose and object geometry. Following Fan et al. (2024), hand pose is evaluated using $\mathrm{MPJPE}_{\mathrm{RA}}$ over 21 MANO keypoints. Prediction and ground truth are independently translated to their wrist roots before comparison; no rotational, scale, or Procrustes alignment is applied.

For object geometry, we uniformly sample $20{,}000$ surface points by triangle area from both the reconstructed and ground-truth canonical meshes. A proper $\mathrm{Sim}(3)$ transformation, consisting of rotation, translation, and uniform scale without reflection, is estimated once per object and kept fixed. CD is the sum of the two mean nearest-surface Euclidean distances and is therefore reported in centimeters rather than squared centimeters. F@5 uses a $5\,\mathrm{mm}$ nearest-surface threshold.

Object trajectory. We evaluate object motion using ATE, ARE, RTE, and RRE following Shin et al. (2026). To avoid dependence on arbitrary canonical object origins, translation trajectories are represented by the area-weighted surface centroids of the posed meshes.

For each demonstration, all manipulated objects share a single fixed-scale rigid alignment between the predicted and ground-truth worlds. The alignment is estimated jointly from all valid object-center correspondences, with each object given equal total weight. No scale adjustment is allowed. This alignment is estimated once per demonstration and is never re-estimated independently for individual objects or frames.

The canonical orientation of a reconstructed object can differ from that of its ground-truth mesh by a constant rotation. We therefore remove one fixed orientation offset using the first common valid frame of each object and keep it unchanged throughout the sequence. No frame-wise rotational alignment is applied. ARE and RRE use the standard geodesic distance on $SO(3)$; object symmetry is not explicitly modeled. RTE and RRE are computed only between consecutive valid prediction frames and never bridge missing frames.

DexImit produces trajectories at a temporal stride of three frames. Its trajectory metrics are therefore evaluated on its native sampled frames, and its relative errors correspond to three-frame intervals; these entries are marked accordingly in the table.

MagicHOI reconstructs its hand-object branches under independent SfM gauges and does not provide a common reconstructed camera trajectory. For MagicHOI only, we use the ground-truth camera trajectory to transform its valid camera-frame object predictions into the ground-truth world before computing the same trajectory metrics. These GT-camera-assisted results are marked with * in the table.

Hand-object interaction. We evaluate relative hand-object geometry using CD_h (Fan et al., 2024). At each valid frame, the predicted and ground-truth posed object surfaces are expressed relative to their corresponding wrist roots in the camera frame, after which the same unsquared bidirectional Chamfer distance as above is computed. No additional per-frame scale, rotation, or rigid alignment is applied.

Physical stability is evaluated using a GraspTTA-style simulation displacement (SD) protocol (Jiang et al., 2021). For each evaluated hand-object configuration, the hand is held static and the object is simulated dynamically in PyBullet for 35 steps at $1/240\,\mathrm{s}$. The same gravity, object mass, contact, solver, and collision settings are used for all methods. SD measures the displacement of the simulated object between its initial and final states.

We additionally report maximum penetration depth $D_{\mathrm{pen}}^{\max}$ (Hasson et al., 2019). For each frame, we measure the largest penetration of MANO vertices into the posed object surface; frames without penetration contribute zero. The reported value is the mean of these frame-wise maxima over the evaluated sequence.

Aggregation and validity. Metrics are aggregated with equal weighting across the relevant hands or objects within each demonstration, followed by equal weighting across demonstrations. Geometry metrics are computed once for each reconstructed canonical object, whereas pose and interaction metrics are evaluated over their valid temporal support. Missing predictions are not interpolated. When a method does not produce the output required by a metric, the corresponding entry is reported as N/A. For methods with incomplete reconstruction coverage, we additionally report the number of valid demonstrations or branches used for evaluation.

D VISUAL SYNTHESIS BASELINE PROTOCOLS

Source-background synthesis. For Kiwi-Edit (Lin et al., 2026), we provide the complete source human-manipulation video together with a reference image showing XHand dexterous hands mounted on Rainbow Robotics RB-Y1 arms. We use the following instruction for all videos:

> "Replace the human hands and arms with XHand dexterous hands mounted on Rainbow Robotics RB-Y1 robot arms, performing the same manipulation task. Keep the original background and all other scene elements unchanged."

The generated video is used directly without additional geometric correction or post-processing.

For GPT-6 Astra (OpenAI, 2026), each source frame is edited using the human frame as the manipulation reference and the same RB-Y1–XHand reference image(s) used to specify the target embodiment. We use the following fixed prompt:

> "Edit the first image. Replace the visible human hand(s) manipulating the object with XHand dexterous hand(s) mounted on a Rainbow Robotics RB-Y1 dual-arm robot, matching the reference image(s). If any wrist, forearm, or arm is visible, replace it with the corresponding RB-Y1 robot arm segment and keep the XHand-arm connection physically plausible. Preserve plausible hand-object contact. Do not add text, extra objects, or a collage layout."

We retain the first generated result for each frame and perform no manual correction or result selection.

Target-background synthesis. For the target-background experiment, GPT-6 Astra receives two visual inputs for each sampled frame: the human source frame and one fixed image of the target real-robot workspace. The source frame specifies the manipulation action, manipulated object, hand-object interaction, and interaction stage, while the fixed target image specifies the camera viewpoint, workspace, background, and overall composition. The robot and object states already present in the target image are not treated as references.

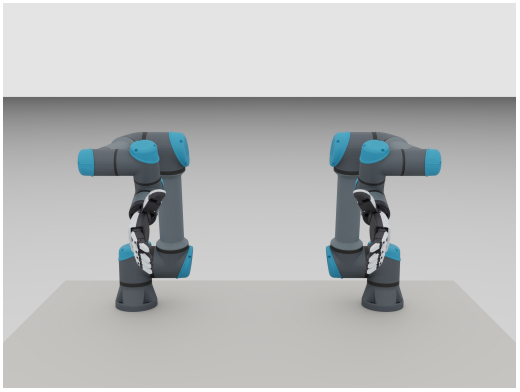
The target embodiment is fixed to two UR5e arms, each equipped with an XHand dexterous hand. We use the same target image and the following prompt for every source frame:

> "Use the current source frame as the reference for the manipulation action, manipulated object, hand-object interaction, and interaction stage. Edit the fixed target workstation image so that a bimanual robot consisting of two XHand dexterous hands, each mounted on a UR5e robot arm, performs the same manipulation shown in the source frame.
>
> Match the manipulated object, its pose, the grasp or contact configuration, and the stage of the manipulation to the source frame as closely as possible. Replace the robot configuration and manipulated-object state already visible in the target image whenever necessary; they do not define the desired action.
>
> Preserve the target image's camera viewpoint, workspace, background, table, overall scene composition, and visual appearance. The generated UR5e arm segments must connect naturally to the XHand hands and remain physically plausible from the target viewpoint.
>
> Do not copy the source image viewpoint or background. Do not preserve the target robot pose when it conflicts with the source manipulation. Do not add text, unrelated objects, extra robot limbs, or a collage layout."

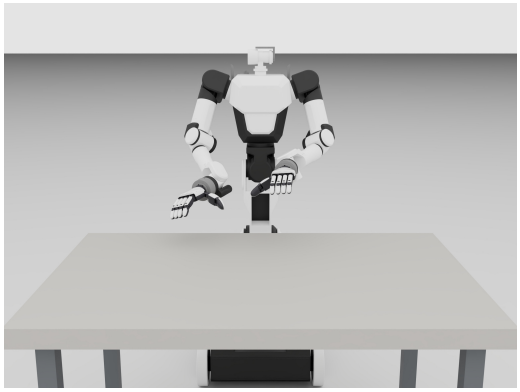
Each frame is generated independently with the same prompt and fixed target image. We use one generation per frame, retain the first result, and apply no manual selection or post-processing. No URDF, CAD, mesh, simulation asset, EgoMatrix output, or additional robot reference is provided to Astra.

E ROBOT DEPLOYMENT

We deploy the transferred trajectories on two real-robot platforms. The first platform consists of two UR5e arms, each equipped with an XHand dexterous hand. The second uses the dual-arm system of a Rainbow Robotics RB-Y1, also equipped with XHand dexterous hands. Fig. 7 shows the two hardware setups used in our experiments.



(a) Dual UR5e arms with XHand dexterous hands.



(b) Rainbow Robotics RB-Y1 with XHand dexterous hands.

Figure 7: **Real-robot platforms used for trajectory replay.** We evaluate transferred trajectories on two embodiments equipped with XHand dexterous hands: a dual-UR5e setup and the dual-arm Rainbow Robotics RB-Y1 platform.

Fig. 8 shows five representative real-world replay sequences. For each task, the upper row contains frames sampled from the source human demonstration, while the lower row shows the corresponding stages of the transferred trajectory replayed on the real robot. The examples include demonstrations collected by us as well as demonstrations from public egocentric datasets.

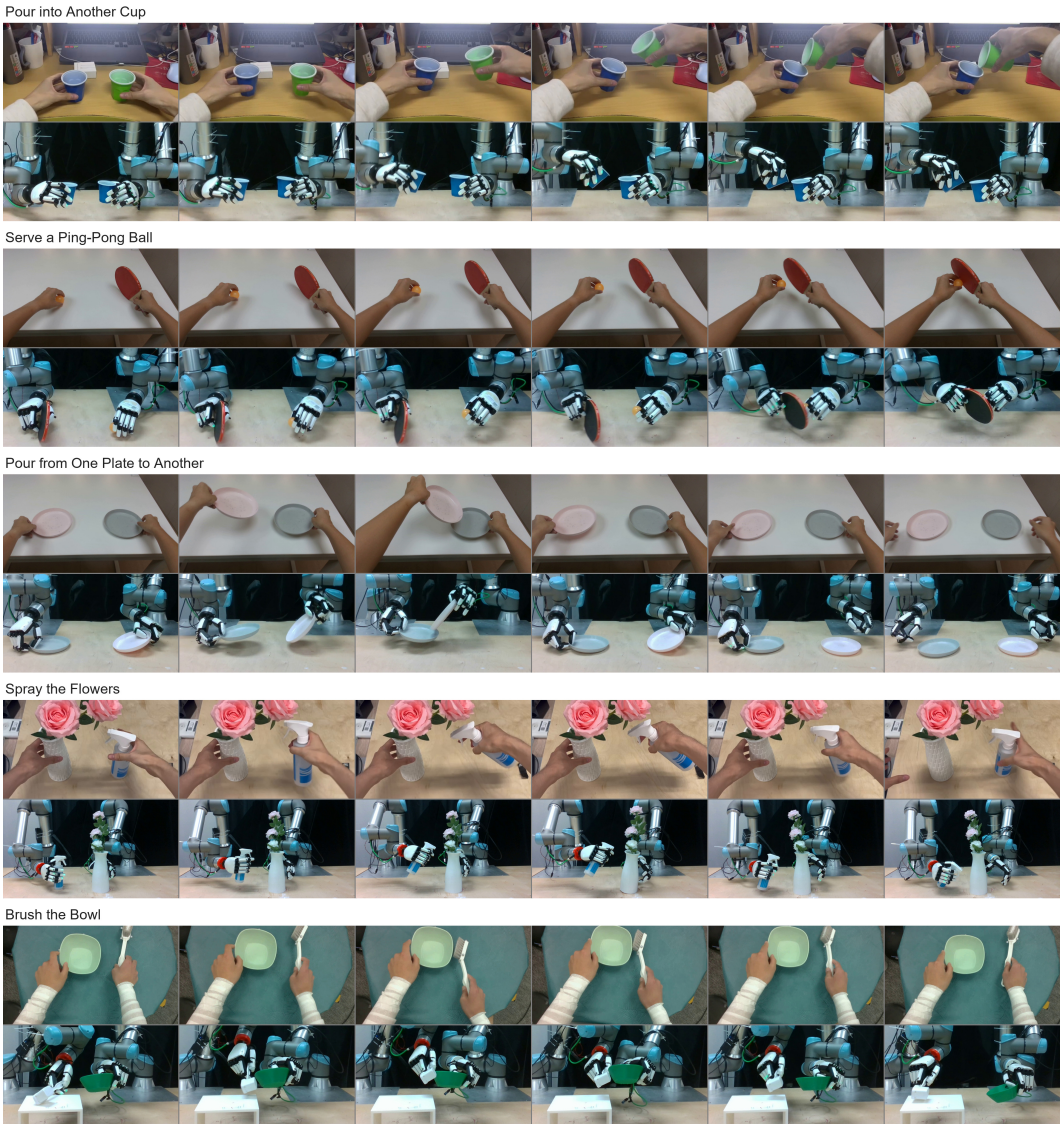


Figure 8: **Real-world trajectory replay across representative tasks.** For each task, the top row shows the source human demonstration and the bottom row shows the corresponding real-robot replay. Time progresses from left to right.