HAND2ROBOT: OBJECT-CENTRIC 4D HAND RECONSTRUCTION FROM HUMAN DEMONSTRATIONS FOR ROBOTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human demonstrations can capture task-relevant manipulation without using the target robot, but the recorded hand motions cannot directly serve as robot action supervision. Hand reconstruction and motion retargeting provide a direct route to such supervision, yet recovering hand motion alone does not ensure physically consistent hand-object interactions. To address this issue, we introduce **Hand2Robot**, an object-centric 4D hand reconstruction framework for robot policy post-training. Given videos of humans performing the target task, Hand2Robot extracts interaction clips, reconstructs object geometry, and generates robot action supervision through hand-object interaction optimization and motion retargeting. For object tracking, we combine observation confidence with temporal motion constraints to select a globally consistent object pose trajectory from per-frame pose candidates. For hand refinement, we use object geometry and motion as a reference and combine staged 2D alignment with object-local contact constraints to refine hand trajectories and finger joint configurations. Experiments on DexYCB and HO3D-v3 show improved object pose accuracy and hand-object interaction quality. Experiments on three real-robot dexterous manipulation tasks further show that reconstructed and retargeted human demonstrations can effectively supplement limited robot data for target-task policy post-training.

1 INTRODUCTION

Deploying pretrained robot policies for dexterous manipulation on a target robot often requires post-training with task-relevant demonstrations. Collecting robot demonstrations at low cost remains challenging, given constraints such as dependence on the target hardware and the need for skilled teleoperators. These constraints motivate the use of human manipulation videos (Grauman et al., 2022; Hoque et al., 2025) as a complementary data source: humans can directly demonstrate task-relevant object interactions without operating the target robot. Using these videos for policy post-training requires translating human motions into the robot's action space while preserving the demonstrated interactions. We therefore investigate how to generate reliable robot-action supervision from human demonstrations to complement limited robot data.

Existing research uses human videos to support robot learning through visual representations, reward signals, or latent action pretraining (Hu et al., 2024; Guzey et al., 2025; Ye et al., 2025a). Another line of work recovers human motion from video and generates robot-action supervision through motion alignment or retargeting (Kareer et al., 2025; Yang et al., 2025). For dexterous manipulation, recovering hand motion also requires preserving the demonstrated hand-object relationships: object geometry constrains feasible hand positions and joint configurations relative to the object, while contact relationships describe how the hand interacts with it (Fan et al., 2024; Xu et al., 2026). Ignoring these relationships can produce visually plausible motion with missing contacts or hand-object interpenetration, compromising the reliability of the generated supervision. Prior work improves manipulation trajectories recovered from video through geometric refinement, contact optimization, or interaction-state constraints (Chen et al., 2026; Wang et al., 2026; Zhang et al., 2026).

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

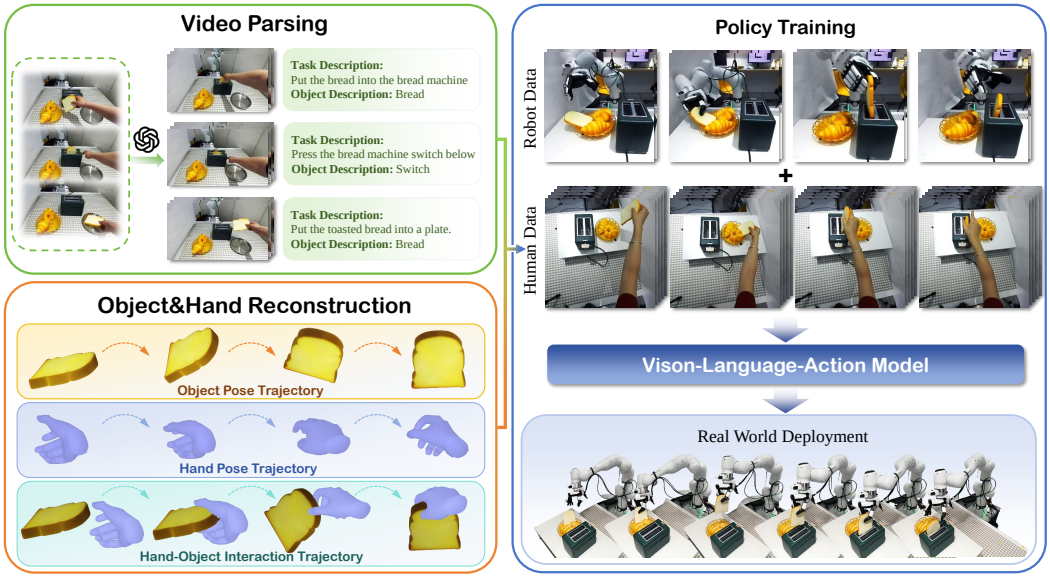


Figure 1: Overview of the Hand2Robot pipeline. Given human demonstration videos, we first parse them into several semantic tasks by VLM. Then we apply our hand-object reconstruction and refinement framework for physically realistic hand trajectories. Finally, these hand trajectories are retargeted to the robot action space, enabling human-robot data co-training on a Vision-Language-Action model for real-world deployment.

However, refining hand motion under geometric and contact constraints still requires reliable estimates of object geometry and pose. Occlusion and pose ambiguity can introduce errors and temporal inconsistencies into object pose estimates (Paliwal et al., 2026), potentially leading hand refinement away from the demonstrated interaction. Moreover, locally plausible contact does not ensure that hand motion agrees with image observations and remains temporally coherent. To address these challenges, we introduce **Hand2Robot**, an object-centric 4D hand reconstruction framework that combines global object trajectory selection with hand refinement constrained by observations and contact. Given RGB-D videos of humans performing the target task, Hand2Robot reconstructs object geometry and combines observational evidence with temporal continuity to select a globally consistent object pose trajectory from per-frame pose candidates. Using the recovered object geometry and motion as references, it then jointly refines the hand's global pose trajectory and finger joint configurations using image alignment together with contact and temporal constraints. Finally, the refined motion is retargeted to the target robot's action space to generate action supervision for policy post-training.

During this reconstruction process, object pose is affected by hand occlusion and object symmetry, while hand motion is affected by partial occlusion and contact ambiguity (Fan et al., 2024; Xu et al., 2026). For object tracking, selecting a single pose independently in each frame can introduce errors under occlusion and accumulate drift, or cause frequent switching between equivalent poses of symmetric objects. Hand2Robot therefore retains multiple pose candidates per frame and combines their observation confidence with temporal continuity between adjacent frames to select a globally consistent object pose trajectory. For hand reconstruction, generic hand estimators often produce incomplete estimates in contact regions and do not use object geometry, so their outputs serve only as initial trajectories (Fan et al., 2024; Xu et al., 2026). Hand2Robot uses contact-aware staged optimization: it first stabilizes hand placement through 2D observation alignment, then introduces contact consistency constraints in the object coordinate frame and penetration penalties to suppress hand-object interpenetration and abrupt changes in contact location.

We evaluate Hand2Robot in terms of reconstruction quality and downstream policy learning. First, on public hand-object interaction benchmarks including DexYCB (Chao et al., 2021) and HO3D-v3 (Hampali et al., 2020), we evaluate object pose, object geometry, hand motion, hand-object inter-

action quality, and temporal stability under a common evaluation protocol, comparing against initialization results and baseline methods. Second, on three real-robot dexterous manipulation tasks, we combine reconstructed and retargeted human demonstrations with limited robot demonstrations for policy post-training, and compare against training with robot demonstrations alone to examine whether the generated human data provides useful action supervision.

Our main contributions are as follows:

- We introduce Hand2Robot, a framework that converts human demonstration videos into action supervision for dexterous robot policy post-training. The framework combines hand-object interaction reconstruction with motion retargeting, enabling human demonstrations to complement limited robot data.
- We propose an object-centric 4D hand reconstruction pipeline that estimates object geometry and pose trajectories, then uses them as fixed references to refine hand motion. By combining confidence-aware temporal object tracking with observation-to-contact hand refinement, our method maintains interaction consistency under occlusion, object symmetry, and contact ambiguity.
- We demonstrate the effectiveness of the generated action supervision for policy post-training on three real-robot dexterous manipulation tasks. The results show that reconstructed and retargeted human demonstrations can effectively complement limited robot data.

2 Related Works

Robot Learning from Human Videos. Human videos offer a scalable complement to robot teleoperation data, but their utility for policy learning depends on whether visual demonstrations can be grounded into action-like supervision. Prior works exploit human videos for representation learning (Ma et al., 2022; Nair et al., 2022; Radosavovic et al., 2023), trajectory or plan extraction (Bharadhwaj et al., 2024; Wen et al., 2023; Wang et al., 2023), cross-embodiment retargeting (Qin et al., 2022; Chen et al., 2022b; 2025b), and action-space learning for vision-language-action models (Ye et al., 2025a; Bu et al., 2025; Bjorck et al., 2025). More recent egocentric approaches recover hand motion or interaction cues to bridge human demonstrations and robot manipulation (Mandikal & Grauman, 2022; Yang et al., 2025; Luo et al., 2025; Bi et al., 2026). However, raw videos usually lack physically grounded hand-object trajectories, making it difficult to mix them directly with robot demonstrations. Our work addresses this gap by reconstructing HOI trajectories from human videos and using the resulting annotations as additional training data for improving policies such as $\pi_{0.5}$ (Intelligence et al., 2025).

Hand-Object Pose Estimation. Recent hand foundation models such as HaMeR and WiLoR provide strong monocular hand mesh initialization (Pavlakos et al., 2024; Potamias et al., 2025), while video methods improve temporal hand trajectories (Yu et al., 2025; Ye et al., 2025b; Zhang et al., 2025). Yet hand-only estimates lack object geometry and contact constraints, causing depth drift, penetration, and unstable contact under occlusion. Hand-object reconstruction methods further introduce object information to constrain the interaction. Template-based or RGB-D/multi-view methods use explicit object geometry and pose (Tekin et al., 2019; Liu et al., 2021; Grady et al., 2021; Yang et al., 2021; Chao et al., 2021; Hampali et al., 2020; Yang et al., 2022), while recent template-free methods use implicit fields, SDFs, NeRFs, diffusion models, or foundation models to recover unknown objects and hand-object configurations (Hasson et al., 2019; Karunratanakul et al., 2020; Chen et al., 2022a; 2023; Fan et al., 2024; Ye et al., 2023; Wang et al., 2025; Liu et al., 2025). These works show that object geometry and contact constraints are essential, but reliable hand-object trajectories remain challenging in human manipulation videos due to occlusion, depth ambiguity, and unstable initialization. Building on these advances, we use foundation models to initialize object and hand states, and introduce two refinement modules: confidence-guided object pose trajectory estimation for the object, and observation-to-contact hand refinement for the hand.

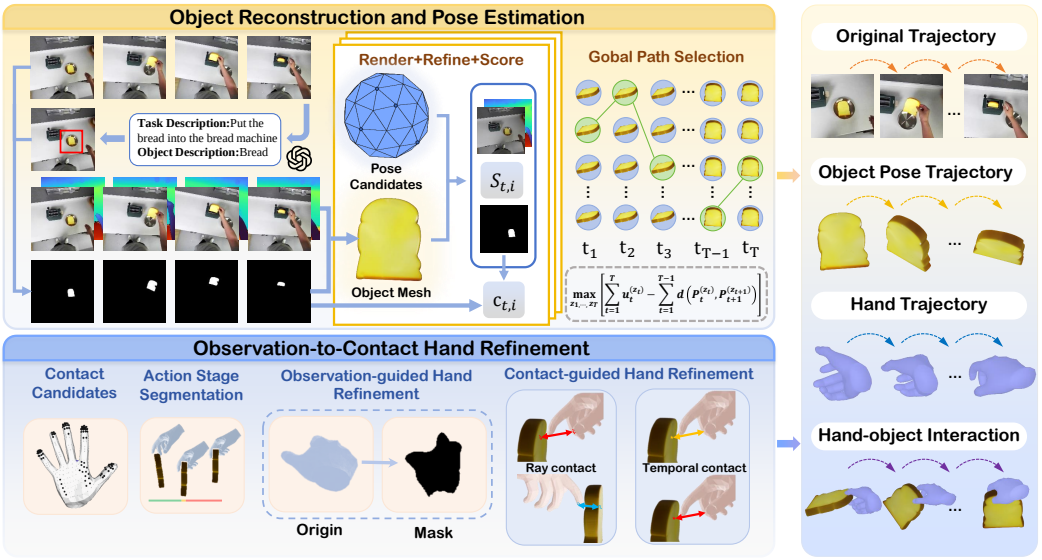


Figure 2: Overview of our object-centric 4D hand reconstruction pipeline.

3 METHOD

We consider the problem of converting human demonstrations into action supervision for dexterous robot policy post-training. The input consists of RGB-D videos of humans performing the target manipulation task, together with the corresponding camera intrinsics. Each video may contain one or more interaction episodes. Hand2Robot reconstructs hand-object interactions and retargets the refined hand motion to the target robot, producing a sequence of robot-action labels temporally aligned with the video observations. The resulting observation-action pairs serve as supplementary training data alongside limited robot demonstrations.

3.1 VIDEO PARSING AND OBJECT GROUNDING

We first use a vision-language model (VLM) (Singh et al., 2025) to segment the input videos into interaction clips and annotate each clip with descriptions of the task and manipulated object. For each clip, we use the object description to guide Rex-Omni (Jiang et al., 2025) in localizing the manipulated object. The resulting detections serve as prompts for SAM3 (Carion et al., 2025), yielding per-frame object masks $\{M_t\}_{t=1}^{T}$ for subsequent object reconstruction and observation-based scoring of pose candidates. We also use MapAnything (Keetha et al., 2025) to recover camera extrinsics for each RGB-D frame.

3.2 OBJECT RECONSTRUCTION AND ROBUST POSE TRACKING

Object Reconstruction. We use SAM 3D (Chen et al., 2025a) to reconstruct the manipulated object as a metric-scale 3D mesh $\mathcal{O}$ from an RGB image and its corresponding object mask. The reconstructed mesh provides a geometric reference for subsequent object pose tracking and hand-object interaction refinement.

Sequential Candidate Selection. Hand occlusion and object symmetry can make per-frame object pose estimates ambiguous. We therefore retain the N refined pose candidates from FoundationPose (Wen et al., 2024) at each frame, together with their ScoreNet scores. This yields the candidate set $\mathcal{P}_t = \{P_t^{(i)}\}_{i=1}^{N}$, where $P_t^{(i)} = (\mathbf{R}_t^{(i)}, \mathbf{t}_t^{(i)})$ denotes the i-th pose candidate at frame t, with raw score $\sigma_t^{(i)}$. We then select an object pose trajectory across the full sequence by combining observation agreement with temporal continuity.

We associate each candidate $P_t^{(i)}$ with a unary weight $u_t^{(i)} \in [0,1]$ that measures its agreement with the visual observation, and define the transition cost between adjacent-frame candidates as:

$$d(P_t^{(i)}, P_{t+1}^{(j)}) = \lambda_{\text{trans}} \|\mathbf{t}_t^{(i)} - \mathbf{t}_{t+1}^{(j)}\|_2 + \lambda_{\text{rot}} \arccos\left(\frac{\text{tr}\left((\mathbf{R}_t^{(i)})^\top \mathbf{R}_{t+1}^{(j)}\right) - 1}{2}\right), \tag{1}$$

where $\lambda_{\text{trans}}, \lambda_{\text{rot}} > 0$ control the translation and rotation smoothness, respectively. The full optimization reads:

$$\max_{z_1,\ldots,z_T} \left[\sum_{t=1}^{T} u_t^{(z_t)} - \sum_{t=1}^{T-1} d(P_t^{(z_t)}, P_{t+1}^{(z_{t+1})})\right]. \tag{2}$$

The optimization in Eq. (2) selects one candidate index $z_t \in \{1,\ldots,N\}$ per frame across all T frames. This induces a graph with T layers and N nodes per layer: node i at layer t carries the unary weight $u_t^{(i)}$, and a directed edge from node i at layer t to node j at layer $t+1$ carries weight $-d\big(P_t^{(i)}, P_{t+1}^{(j)}\big)$. Since all edges point strictly from t to $t+1$, the graph is acyclic and layered, reducing Eq. (2) to a longest-path problem solvable by Viterbi decoding (Rabiner, 1989) in $O(TN^2)$.

Construction of Unary Weights. We use the ScoreNet scores $\{\sigma_t^{(i)}\}$ to rank pose candidates within each frame and apply per-frame min-max normalization:

$$\bar{\sigma}_t^{(i)} = \frac{\sigma_t^{(i)} - \min_j \sigma_t^{(j)}}{\max_j \sigma_t^{(j)} - \min_j \sigma_t^{(j)}}, \tag{3}$$

This normalization preserves the within-frame ranking, but the highest-scoring candidate receives a value of 1 even when all candidates are inaccurate. We therefore supplement the normalized scores with image-space agreement. For each candidate $P_t^{(i)}$, we render the object mask and compute its IoU with the observed mask M_t:

$$c_t^{(i)} = \text{IoU}\left(\mathcal{R}_{\text{mask}}(\mathcal{O}, P_t^{(i)}), M_t\right), \tag{4}$$

The unary weight combines the normalized ScoreNet score with this observation-agreement weight:

$$u_t^{(i)} = c_t^{(i)} \cdot \bar{\sigma}_t^{(i)}. \tag{5}$$

This weighting reduces the contribution of candidates whose rendered masks disagree with the observation. When mask agreement is low for all candidates in a frame, the reduced unary weights increase the relative influence of temporal transition costs in Eq. (2) on trajectory selection.

3.3 Observation-to-Contact Hand Refinement

We obtain per-frame MANO estimates (Romero et al., 2017) from WiLoR (Potamias et al., 2025) and use tracking and depth-based alignment for metric initialization. Using the reconstructed object geometry and pose trajectory as fixed references, we jointly refine the hand's global pose trajectory and finger joint configurations across the sequence. The refinement emphasizes 2D observation alignment before reliable contact is available and incorporates contact constraints during interaction.

Action stage segmentation. We segment each action into grasping and interaction stages using object motion and hand–object proximity. Let T_t^o be the object translation and d_t^{ho} the minimum hand–object distance:

$$e_t = \alpha \frac{\|T_t^o - T_{t-1}^o\|_2}{\delta_{\text{trans}}} + \beta \max\left(0, 1 - \frac{d_t^{\text{ho}}}{\delta_{\text{ho}}}\right), \tag{6}$$

where α, β are cue weights and $\delta_{\text{trans}}, \delta_{\text{ho}}$ are normalization thresholds. We select the earliest frame satisfying $e_t > \delta_{\text{on}}$ as the anchor t_a. Single-frame refinement at t_a yields a hand translation offset $\Delta\tau = \tau_a^\star - \tau_a^0$, where τ_a^0 is the initial hand translation at the anchor frame and $\tau_a^\star$ is the refined translation after single-frame optimization. This offset initializes the entire segment.

We then optimize the MANO parameters $\Theta_t = \{s_t, \tau_t, \omega_t, \theta_t, \beta_t\}$ (see Appendix A.3) while keeping the object pose $\mathcal{O}_t = (R_t^o, T_t^o)$ fixed:

$$\mathcal{L} = \lambda_{\text{2d}}\mathcal{L}_{\text{2d}} + \lambda_{\text{match}}\mathcal{L}_{\text{match}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{sdf}}\mathcal{L}_{\text{sdf}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{7}$$

Here, $\mathcal{L}_{\text{2d}}$ aligns masks, $\mathcal{L}_{\text{match}}$ encourages per-frame contact matching, $\mathcal{L}_{\text{temp}}$ promotes temporal contact consistency, $\mathcal{L}_{\text{sdf}}$ penalizes penetration, and $\mathcal{L}_{\text{sm}} + \mathcal{L}_{\text{reg}}$ regularize the trajectory. All $\lambda_{(\cdot)}$ are scalar weights.

2D-guided hand alignment. During the grasping stage, we emphasize 2D alignment before reliable contact constraints are available. We use the stage-weighted mask loss $\mathcal{L}_{\text{2d}} = \frac{1}{B}\sum_t \gamma_t \, \text{SoftIoU}(\hat{M}_h^t, M_h^t)$, where B is the number of optimized frames, $\hat{M}_h^t$ and M_h^t are the rendered and observed hand masks, and SoftIoU is a differentiable soft IoU loss. We set $\gamma_t = \gamma_g > 1$ for grasping frames and $\gamma_t = 1$ for interaction frames.

Per-frame contact matching. During interaction, we adopt ray-based 2D-to-3D contact matching (Lin et al., 2025). For each active MANO contact candidate $v_{t,k}$, a camera ray through its image projection intersects the posed object mesh at $o_{t,k}$. We minimize the matching loss

$$\mathcal{L}_{\text{match}} = \frac{1}{|\mathcal{I}_{\text{inter}}|} \sum_{t \in \mathcal{I}_{\text{inter}}} \frac{1}{\max(|\mathcal{C}_t|, 1)} \sum_{k \in \mathcal{C}_t} \|v_{t,k} - o_{t,k}\|_2, \tag{8}$$

where $\mathcal{I}_{\text{inter}}$ is the interaction-frame set and $\mathcal{C}_t$ is the active contact set.

Temporal contact stabilization. Per-frame matching can assign the same hand candidate to different object regions in adjacent frames. To improve temporal consistency, we regularize contact locations in the object coordinate frame, accounting for the object's rigid motion. We transform the ray-hit point and hand candidate as $\bar{o}_{t,k} = (R_t^o)^\top (o_{t,k} - T_t^o)$ and $\bar{v}_{t,k} = (R_t^o)^\top (v_{t,k} - T_t^o)$. For candidates active in two adjacent frames, $\mathcal{C}_{t,t+1} = \mathcal{C}_t \cap \mathcal{C}_{t+1}$, we use the average ray-hit location as a stop-gradient local anchor, $a_{t,k} = \frac{1}{2}(\text{sg}(\bar{o}_{t,k}) + \text{sg}(\bar{o}_{t+1,k}))$, and minimize

$$\mathcal{L}_{\text{temp}} = \sum_{t,\, t+1 \in \mathcal{I}_{\text{inter}}} \sum_{k \in \mathcal{C}_{t,t+1}} \left[\rho(\|\bar{v}_{t,k} - a_{t,k}\|_2) + \rho(\|\bar{v}_{t+1,k} - a_{t,k}\|_2)\right]. \tag{9}$$

Here sg is stop-gradient and $\rho(x) = \sqrt{x^2 + \epsilon}$ is the Charbonnier robust loss with a small $\epsilon > 0$ for numerical stability. This term encourages the same hand candidate to remain near a common local contact region across adjacent frames, reducing inconsistent contact drift.

After warmup, we enable $\mathcal{L}_{\text{sdf}}$ to penalize hand–object penetration, while $\mathcal{L}_{\text{sm}}$ suppresses abrupt motion changes and $\mathcal{L}_{\text{reg}}$ limits deviations from the initialization. Further details are provided in Appendix A.3.

3.4 From Human Hand to Dexterous Hand

The final step converts the refined MANO motion into robot action supervision. We use GeoRT (Yin et al., 2025) to obtain joint configurations for the target dexterous hand. The global hand pose trajectory is first transformed from camera to world coordinates using the camera extrinsics from Sec. 3.1. We then use representative initial object poses in the human and robot demonstrations to estimate a world-to-robot transformation and map the trajectory into the robot frame. The aligned poses and retargeted joint configurations are encoded using the same action representation as robot demonstrations and paired with the corresponding video observations for policy post-training.

4 Experimental Results

Our experiments evaluate Hand2Robot along two axes. First, we evaluate whether the proposed object-centric reconstruction improves hand-object trajectory quality on public benchmarks. Second, we evaluate whether the reconstructed and retargeted human demonstrations improve policy learning in real-robot dexterous manipulation tasks.

4.1 Hand-Object Reconstruction Evaluation

We evaluate Hand2Robot in terms of hand-object reconstruction quality on public benchmarks and downstream policy learning on real-robot dexterous manipulation tasks.

Datasets and Metrics. We evaluate on DexYCB (Chao et al., 2021) and HO3D-v3 (Hampali et al., 2020). For DexYCB, we select a subset of 1,000 sequences for object pose evaluation, from which

we further sample 120 sequences for hand evaluation. HO3D-v3 contains longer manipulation sequences. We sample a subset of 12 sequences for evaluation. For object pose, we report translation error, rotation error, and ADD/ADD-S AUC (Xiang et al., 2017). We evaluate hand pose accuracy (e.g., MPJPE, MRPE), physical interaction metrics (e.g., penetration volume, IV), and jerk magnitude (JM) to compare Hand2Robot against the WiLoR baseline. Metric definitions and alignment procedures are described in Appendix C.

Object Reconstruction Quality. We additionally evaluate object reconstruction on DexYCB. Table 5 in Appendix D.1 reports CD, F@5, and F@10 for HOLD, MagicHOI, and our pipeline. Our pipeline achieves a CD of 0.57 cm^2, with F@5 and F@10 of 49.35% and 87.54%, respectively.

Object Pose Tracking Results. Table 1 reports object pose tracking results on DexYCB. Compared with the default FoundationPose output, our global pose-path selection reduces translation error from 6.80 to 3.44 cm and rotation error from 27.14° to 18.47°. ADD AUC increases from 98.96 to 99.48, while ADD-S AUC increases from 99.15 to 99.62. These results show improved object pose accuracy when selecting the trajectory using both observation scores and temporal constraints.

Table 1: Quantitative object pose results on DexYCB using reconstructed meshes and first-frame alignment. See Appendix C for evaluation details.

Method	Translation error (cm) ↓	Rotation error (°) ↓	ADD AUC ↑	ADD-S AUC ↑
FoundationPose (Wen et al., 2024)	6.80	27.14	98.96	99.15
Ours w/o d, c	4.88	61.63	99.07	99.44
Ours w/o d	3.62	38.20	99.37	99.61
Ours w/o c	**3.44**	19.86	99.47	**99.62**
Ours	**3.44**	**18.47**	**99.48**	**99.62**

Table 2: Quantitative evaluation for in-scene object manipulation. We evaluate on HO3D-v3 and DexYCB, comparing the WiLoR initialization with our refined hand trajectories. We highlight the best values in bold.

Dataset	**Method**	MPJPE ↓	MRPE ↓	IV-mean ↓	IV-max ↓	JM ↓
DexYCB (Chao et al., 2021)	WiLoR (Potamias et al., 2025)	8.35	97.31	2.57	5.41	2.58
	Ours	**7.58**	**38.49**	**1.07**	**3.25**	**2.49**
HO3D-v3 (Hampali et al., 2020)	WiLoR	**4.22**	24.22	20.11	64.82	0.86
	Ours	4.40	**16.99**	**3.42**	**22.40**	**0.34**

Hand Pose and Trajectory Refinement Results. Table 2 reports results on DexYCB (Chao et al., 2021) and HO3D-v3 (Hampali et al., 2020). While WiLoR provides strong per-frame initialization, it lacks temporal and contact constraints, leading to trajectory drift and object penetration. By integrating 2D mask alignment, ray-based contact, SDF, and temporal smoothness, our optimization yields temporally consistent and physically plausible trajectories. Specifically, our refinement reduces wrist trajectory error (MRPE) on DexYCB and hand-object interpenetration (IV) on both datasets. These results show reduced hand-object interpenetration and improved wrist trajectory accuracy and smoothness.

4.2 Real-Robot Experiments

4.2.1 Experimental Setup and Evaluation Protocol

We evaluate whether the action supervision generated by Hand2Robot supports policy post-training for target tasks on a real robot. Our platform consists of a Franka arm and an xHand dexterous hand. Human and robot demonstrations are recorded using the same fixed side-view camera system. Human RGB-D demonstrations are converted into action supervision in the target robot coordinate frame through hand-object interaction reconstruction and motion retargeting, and are combined with limited robot demonstrations for $\pi_{0.5}$ post-training. This setup allows us to examine how the processing of human demonstrations affects downstream policy performance.

Our main comparisons cover three real-robot tasks shown in Figure 3: Place Vase, Open Box, and Insert Toast, which assess policy performance in pick-and-place, opening, and insertion, respectively.

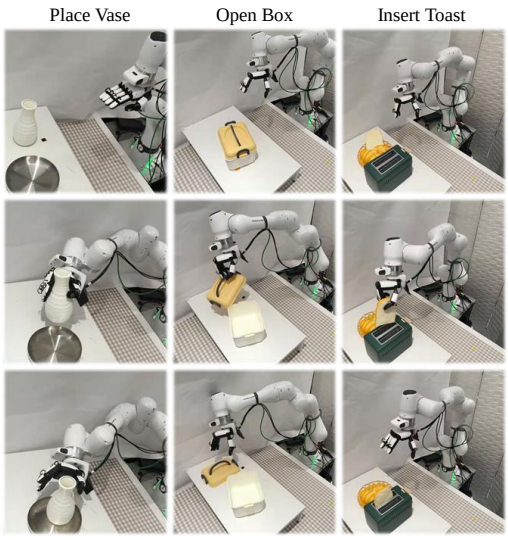


Figure 3: Examples of real-robot manipulation tasks. Each column shows one task, with the three rows corresponding to its initial, intermediate, and successful final states. The tasks are Place Vase, Open Box, and Insert Toast, from left to right.

All conditions use the same policy architecture and training procedure. In the main comparison of action generation methods, we hold the 20 robot demonstrations and 100 source human videos per task fixed, while varying how human action supervision is generated. The separate budget study varies the number of demonstrations. During mixed-data post-training, human and robot training samples are drawn at a 1:1 ratio. Further details are provided in Appendix F.

We use task success rate as the primary evaluation metric. Each trained policy is evaluated in 20 trials per task, with initial object positions randomly sampled from the training distribution. Success rate is the number of successful trials divided by 20. We report per-task success rates and their unweighted mean.

4.2.2 Downstream Effects of Human Action Supervision

We investigate how different methods for generating action supervision from the same human demonstrations affect downstream policy performance. We process the same 100 human demonstrations using WiLoR with basic retargeting, MagicHOI, and Hand2Robot, and combine each resulting dataset with the same 20 robot demonstrations for $\pi_{0.5}$ post-training. For MagicHOI, we use its hand optimization module within a common surrounding pipeline. A policy trained with only the 20 robot demonstrations serves as a reference.

Table 3: Task success rates (%) after policy post-training with different sources of human action supervision. R and H denote the numbers of robot and human demonstrations per task. The mixed-data conditions use the same 20 robot demonstrations and 100 source human videos. Mean is the unweighted average across the three tasks.

Training data and method	Place Vase	Open Box	Insert Toast	Mean
20R	45.0	60.0	40.0	48.3
20R + 100H, WiLoR	25.0	20.0	10.0	18.3
20R + 100H, MagicHOI	40.0	25.0	15.0	26.7
20R + 100H, Hand2Robot	**80.0**	**85.0**	**75.0**	**80.0**

As shown in Table 3, Hand2Robot achieves a mean success rate of 80.0% across the three tasks, compared with 18.3% for WiLoR with basic retargeting and 26.7% for MagicHOI. Hand2Robot achieves higher success rates on all three tasks. With the source human videos and robot demonstrations held fixed, these results show that Hand2Robot generates more effective action supervision for policy post-training than the compared methods.

Compared with training on the 20 robot demonstrations alone, adding 100 human demonstrations processed by Hand2Robot increases the mean success rate from 48.3% to 80.0%. The gains extend across all three tasks, with success rates improving by 35, 25, and 35 percentage points on Place Vase, Open Box, and Insert Toast, respectively. With the robot demonstration subset held fixed, these improvements show that human demonstrations processed through our reconstruction and retargeting pipeline provide effective additional action supervision, supporting target-task policy post-training with limited robot data.

4.2.3 Robot Demonstration Budget and the Value of Human Data

We further compare 20R+100H with robot-only budgets of 20R, 40R, and 60R to assess whether Hand2Robot-generated supervision can reduce robot demonstration requirements.

Table 4: Success rates (%) under different demonstration budgets and action-supervision sources. R and H denote the numbers of robot and human demonstrations per task, respectively. Do As I Do includes Sharpa-to-xHand adaptation after dynamics-based post-optimization. Mean is the unweighted average across the three tasks. Conditions shared with Table 3 use the same results.

Method	Robot	Human	Place Vase	Open Box	Insert Toast	Mean
Robot-only	20	0	45.0	60.0	40.0	48.3
Robot-only	40	0	60.0	65.0	55.0	60.0
Robot-only	60	0	80.0	90.0	75.0	81.7
Do As I Do (Paliwal et al., 2026)	20	50	50.0	45.0	40.0	45.0
Hand2Robot	20	50	50.0	70.0	45.0	55.0
Hand2Robot	20	100	80.0	85.0	75.0	80.0
Hand2Robot	0	100	15.0	10.0	0.0	8.3

As shown in Table 4, training with 20R+100H achieves a mean success rate of 80.0%, approaching the 81.7% obtained with 60R and exceeding the 48.3% and 60.0% obtained with 20R and 40R, respectively. At the task level, 20R+100H matches the success rates of 60R on Place Vase and Insert Toast, and is five percentage points lower on Open Box. These results demonstrate the value of Hand2Robot-generated supervision for achieving high task success rates with a smaller robot demonstration budget.

We further examine the effect of human demonstration quantity while fixing the robot demonstration budget at 20. Increasing the number of human demonstrations from 0 to 50 and 100 raises the mean success rate from 48.3% to 55.0% and 80.0%, respectively. Within the evaluated range, additional human demonstrations processed by Hand2Robot provide further benefits to policy post-training. We additionally evaluate Do As I Do (Paliwal et al., 2026), retaining its dynamics-based post-optimization and adapting its Sharpa outputs to xHand. At 20R+50H, this adapted baseline achieves success rates of 50.0%, 45.0%, and 40.0% on Place Vase, Open Box, and Insert Toast, respectively, with a mean of 45.0%, compared with 55.0% for Hand2Robot under the same demonstration budget. Training with 100H alone achieves a mean success rate of 8.3%, compared with 80.0% for 20R+100H, highlighting the benefit of combining human and robot demonstrations in this setting.

5 Conclusion

We presented Hand2Robot, a framework for generating action supervision from human demonstration videos for dexterous robot policy post-training. It combines global object pose trajectory selection with refinement of hand trajectories and joint configurations under geometric and contact constraints, followed by motion retargeting. Experiments on DexYCB and HO3D-v3 show improved object pose accuracy and hand-object interaction quality. Three real-robot tasks further demonstrate that the generated supervision complements limited robot data. Future work will address deformable objects and more dynamic contact-rich manipulation.

AI USE STATEMENT

We used generative AI tools to provide feedback on experimental design, assist with analyzing changes in evaluation metrics, and support data processing and qualitative analysis. We also used these tools for literature search and synthesis, language editing, manuscript organization, and the development of supporting code and research artifacts. The authors reviewed and revised the AI-assisted manuscript text to ensure consistency with the research methods, experimental results, and supporting evidence. The authors take full responsibility for the final content of this work.

REFERENCES

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*, pp. 306–324. Springer, 2024.

Hongzhe Bi, Lingxuan Wu, Tianwei Lin, Hengkai Tan, Zhizhong Su, Hang Su, and Jun Zhu. H-rdt: Human manipulation enhanced bimanual robotic manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 18135–18143, 2026.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Qingwen Bu, Yanting Yang, Jisong Cai, Shenyuan Gao, Guanghui Ren, Maoqing Yao, Ping Luo, and Hongyang Li. Univla: Learning to act anywhere with task-centric latent actions. *arXiv preprint arXiv:2505.06111*, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, 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, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9044–9053, 2021.

Hongyi Chen, Tony Dong, Tiancheng Wu, Liquan Wang, Yash Jangir, Yaru Niu, Yufei Ye, Homanga Bharadhwaj, Zackory Erickson, and Jeffrey Ichnowski. Dexterous manipulation policies from rgb human videos via 3d hand-object trajectory reconstruction. *arXiv preprint arXiv:2602.09013*, 2026.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025a.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European conference on computer vision*, pp. 231–248. Springer, 2022a.

Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12890–12900, 2023.

Zerui Chen, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Vividex: Learning vision-based dexterous manipulation from human videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3336–3343. IEEE, 2025b.

Zoey Qiuyu Chen, Karl Van Wyk, Yu-Wei Chao, Wei Yang, Arsalan Mousavian, Abhishek Gupta, and Dieter Fox. Dextransfer: Real world multi-fingered dexterous grasping with minimal human demonstrations. *arXiv preprint arXiv:2209.14284*, 2022b.

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*, pp. 494–504, 2024.

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*, pp. 1471–1481, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Irmak Guzey, Yinlong Dai, Georgy Savva, Raunaq Bhirangi, and Lerrel Pinto. Bridging the human to robot dexterity gap through object-oriented rewards. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3344–3351. IEEE, 2025. URL https://object-rewards.github.io/.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3196–3206, 2020.

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*, pp. 11807–11816, 2019.

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.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. *arXiv preprint arXiv:2510.12798*, 2025.

Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13226–13233. IEEE, 2025.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV)*, pp. 333–344. IEEE, 2020.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.

Dixuan Lin, Tianyou Wang, Zhuoyang Pan, Yufu Wang, Lingjie Liu, and Kostas Daniilidis. Zero-shot reconstruction of in-scene object manipulation from video. *arXiv preprint arXiv:2512.19684*, 2025.

Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14687–14697, 2021.

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

Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: vision-language-action pretraining from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 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. *arXiv preprint arXiv:2210.00030*, 2022.

Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pp. 651–661. PMLR, 2022.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 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.

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*, 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 Computer Vision and Pattern Recognition Conference*, pp. 12242–12254, 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*, pp. 570–587. Springer, 2022.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Conference on Robot Learning*, pp. 416–426. PMLR, 2023.

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), November 2017.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4511–4520, 2019.

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*, pp. 201–221. PMLR, 2023.

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

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*, pp. 5957–5968, 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*, pp. 17868–17879, 2024.

Chuan Wen, Xingyu Lin, John So, Kai Chen, Qi Dou, Yang Gao, and Pieter Abbeel. Any-point trajectory modeling for policy learning. *arXiv preprint arXiv:2401.00025*, 2023.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

Hao Xu, Yilin Liu, Yinqiao Wang, Chi-Wing Fu, and Niloy J Mitra. Choir: Contact-aware 4d hand-object interaction reconstruction. *arXiv preprint arXiv:2605.20992*, 2026.

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

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20953–20962, 2022.

Ruihan Yang, Qinxi Yu, Yecheng Wu, Rui Yan, Borui Li, An-Chieh Cheng, Xueyan Zou, Yunhao Fang, Xuxin Cheng, Ri-Zhao Qiu, et al. Egovla: Learning vision-language-action models from egocentric human videos. *arXiv preprint arXiv:2507.12440*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Se June Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. In *International Conference on Learning Representations*, volume 2025, pp. 28213–28239, 2025a.

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*, pp. 19717–19728, 2023.

Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J Black. Predicting 4d hand trajectory from monocular videos. *arXiv preprint arXiv:2501.08329*, 2025b.

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 *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 17376–17382. IEEE, 2025.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 27716–27726, 2025.

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

Yuhong Zhang, Zihan Gao, Shengpeng Li, Ling-Hao Chen, Kaisheng Liu, Runqing Cheng, Xiao Lin, Junjia Liu, Zhuoheng Li, Jingyi Feng, Ziyan He, Jintian Lin, Zheyan Huang, Zhifang Liu, and Haoqian Wang. RoboWheel: A data engine from real-world human demonstrations for cross-embodiment robotic learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6664–6674, June 2026.

APPENDIX

A METHOD DETAILS

A.1 PROMPT FOR VLM-BASED SEGMENTATION AND ANNOTATION

Due to its length, we do not include the full prompt inline in the appendix, as doing so would substantially reduce readability. Instead, we provide the complete prompt template as a separate text file, `prompt.txt`, in the supplementary material. The file contains the exact prompt used for segmenting human demonstration videos into single-interaction task clips and annotating each clip with task descriptions, directly contacted objects, hand labels, and concise interaction summaries.

A.2 INTERPRETATION OF IOU-BASED POSE WEIGHTS

The IoU term measures agreement between the rendered object silhouette and the observed mask. High IoU indicates close silhouette agreement, although distinct poses can produce similar projections, particularly for symmetric objects. Conversely, partial visibility and segmentation errors can reduce IoU even for an accurate pose. We therefore combine this observation-agreement weight with ScoreNet scores and temporal transition costs when selecting the object trajectory.

A.3 OBSERVATION-TO-CONTACT HAND REFINEMENT

This section provides implementation details for the hand refinement module described in the main paper. Given per-frame MANO initialization from WiLoR (Potamias et al., 2025) and a fixed object trajectory, we optimize the hand trajectory and pose while keeping the object mesh and object poses fixed.

Inputs and Parameterization

For each video segment, the module takes as input the initial MANO hand parameters, hand and object masks, camera intrinsics, a reconstructed object mesh $\Omega = (\mathcal{V}_o, \mathcal{F}_o)$, an object SDF, and the object poses $\mathcal{O}_t = (R_t^o, T_t^o)$. The hand at frame t is represented by

$$\Theta_t = \{s_t, \tau_t, \omega_t, \theta_t, \beta_t\}, \tag{10}$$

where s_t is the global hand scale, τ_t is the hand translation, ω_t is the global orientation, θ_t is the MANO articulation parameter, and β_t is the MANO shape parameter. During hand refinement, we optimize s_t together with the other MANO parameters, while keeping the metric-scale object mesh from SAM 3D and its pose trajectory fixed.

Action Stage Segmentation and Anchor Initialization

We identify the transition from grasping to interaction using object motion and hand–object proximity. Object poses are first projected to valid $SE(3)$ transforms and lightly smoothed. Let T_t^o be the object translation and d_t^{ho} the minimum hand–object distance. The interaction score is

$$e_t = \alpha \frac{\|T_t^o - T_{t-1}^o\|_2}{\delta_{\text{trans}}} + \beta \max\left(0, 1 - \frac{d_t^{\text{ho}}}{\delta_{\text{ho}}}\right), \tag{11}$$

where α, β are cue weights and $\delta_{\text{trans}}, \delta_{\text{ho}}$ are normalization thresholds. We select the earliest frame satisfying $e_t > \delta_{\text{on}}$ as the interaction onset and anchor t_a. Frames before t_a are labeled as grasping, and frames from t_a onward are labeled as interaction.

Before batch optimization, we optimize the anchor frame with the single-frame optimizer and compute a segment-level translation offset $\Delta\tau = \tau_a^\star - \tau_a^0$, where τ_a^0 and $\tau_a^\star$ are the initial and refined translations at the anchor. This offset is added to all keyframes in the segment.

Batch Objective

After anchor initialization, keyframes in the same action segment are optimized jointly. The objective is

$$\mathcal{L} = \lambda_{\text{2d}}\mathcal{L}_{\text{2d}} + \lambda_{\text{match}}\mathcal{L}_{\text{match}} + \lambda_{\text{temp}}\mathcal{L}_{\text{temp}} + \lambda_{\text{sdf}}\mathcal{L}_{\text{sdf}} + \lambda_{\text{sm}}\mathcal{L}_{\text{sm}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}. \tag{12}$$

Here $\mathcal{L}_{\text{2d}}$ aligns hand masks, $\mathcal{L}_{\text{match}}$ enforces per-frame ray contact, $\mathcal{L}_{\text{temp}}$ stabilizes contact over time, $\mathcal{L}_{\text{sdf}}$ penalizes penetration and encourages surface contact, and $\mathcal{L}_{\text{sm}}$ and $\mathcal{L}_{\text{reg}}$ regularize temporal smoothness and deviation from the initialization.

Stage-Weighted 2D Mask Alignment

During grasping, contact is not yet available, so image evidence is the main constraint. We render the hand mask $\hat{M}_h^t$ and compare it with the observed hand mask M_h^t using a soft IoU loss:

$$\mathcal{L}_{\text{2d}} = \frac{1}{B}\sum_{t=1}^{B}\gamma_t\left(1 - \frac{\langle \hat{M}_h^t, M_h^t\rangle + \epsilon}{\|\hat{M}_h^t\|_1 + \|M_h^t\|_1 - \langle \hat{M}_h^t, M_h^t\rangle + \epsilon}\right), \tag{13}$$

where B is the number of optimized frames, ϵ stabilizes the division, and $\gamma_t = \gamma_g > 1$ for grasping frames and 1 otherwise.

Ray-Based Contact Matching.

During interaction, the hand should be constrained by object geometry. We follow the ray-based 2D-to-3D contact construction from REISOM (Lin et al., 2025). Let $\mathcal{Z}$ be the MANO contact-zone vertex set, and let $v_{t,k}$ denote the 3D position of the k-th contact vertex at frame t. A contact candidate is active if its 2D projection falls inside the object mask:

$$\mathcal{C}_t = \{k \in \mathcal{Z} \mid \Pi_K(v_{t,k}) \in M_o^t\}. \tag{14}$$

For each active candidate, we cast a camera ray through the projected pixel and intersect it with the posed object mesh Ω_t to obtain an object-side surface point $o_{t,k}$. The matching loss is

$$\mathcal{L}_{\text{match}} = \frac{1}{|\mathcal{I}_{\text{inter}}|}\sum_{t\in\mathcal{I}_{\text{inter}}}\frac{1}{\max(|\mathcal{C}_t|,1)}\sum_{k\in\mathcal{C}_t}\|v_{t,k} - o_{t,k}\|_2. \tag{15}$$

This loss provides a 3D contact target supported jointly by the object mask and the current object pose.

Object-Local Temporal Contact Stabilization.

Per-frame ray matching may assign the same hand candidate to different object regions over time. We therefore compare contact points in the object coordinate frame:

$$\bar{o}_{t,k} = (R_t^o)^\top(o_{t,k} - T_t^o), \qquad \bar{v}_{t,k} = (R_t^o)^\top(v_{t,k} - T_t^o). \tag{16}$$

For candidates active in adjacent interaction frames, $\mathcal{C}_{t,t+1} = \mathcal{C}_t \cap \mathcal{C}_{t+1}$, we define a stop-gradient local anchor

$$a_{t,k} = \frac{1}{2}\left(\text{sg}(\bar{o}_{t,k}) + \text{sg}(\bar{o}_{t+1,k})\right), \tag{17}$$

and minimize

$$\mathcal{L}_{\text{temp}} = \sum_{t:\, t,t+1\in\mathcal{I}_{\text{inter}}}\sum_{k\in\mathcal{C}_{t,t+1}}\left[\rho(\|\bar{v}_{t,k} - a_{t,k}\|_2) + \rho(\|\bar{v}_{t+1,k} - a_{t,k}\|_2)\right], \tag{18}$$

where sg denotes stop-gradient and $\rho(x) = \sqrt{x^2+\epsilon}$ is a smooth distance penalty with a small constant ϵ. In implementation, we discard non-finite pairs and pairs with large object-local distances to avoid unstable ray correspondences.

SDF Warmup, Smoothness, and Regularization

SDF contact and penetration losses are enabled only after a warmup stage. This prevents early SDF gradients from conflicting with image alignment and ray contact when the initial hand is still far from the object. After warmup, the SDF term removes penetration and improves surface-level contact. The smoothness term is computed over keyframes and penalizes high-frequency changes in hand pose, orientation, and translation. We use second-order smoothness by default, which suppresses acceleration while allowing approximately constant-velocity motion. The regularization term keeps scale, translation, pose, and shape close to the initial foundation-model prediction.

A.4 World-to-Robot Coordinate Alignment

After retargeting the reconstructed human hand motion, the resulting trajectories are still expressed in the camera coordinate system. We first transform all reconstructed hand and object trajectories into a shared world coordinate frame using the estimated camera extrinsics. To obtain the world-to-robot coordinate transformation, we estimate a representative initial object pose for the human and robot demonstrations, respectively. For the human demonstrations, we collect the first-frame object pose of each valid sequence in the world frame. The representative translation is the mean of a Gaussian fitted to these initial positions, while the representative rotation is computed by averaging on the rotation manifold. For the robot demonstrations, we estimate object poses using the camera in the robot setup and transform them into the robot coordinate frame through hand-eye calibration, then compute the corresponding representative initial object pose. The rigid transformation mapping the human-domain representative pose to the robot-domain representative pose is used as the world-to-robot alignment matrix. Applying this transformation expresses the global poses of the reconstructed and retargeted human trajectories in the same coordinate system as the robot demonstrations.

B Hyperparameters Selection

B.1 Hyperparameters for Object Pose Tracking

In the object pose estimation stage, we set the number of pose candidates to ($N = 252$). Following FoundationPose, these candidates are generated by uniformly sampling object viewpoints and in-plane rotations, thereby providing broad coverage of the rotation space before pose refinement and scoring. For the temporal candidate selection objective, we set ($\lambda_{\text{trans}} = 1$) and ($\lambda_{\text{rot}} = 10$). These weights are not task-specific tuned parameters; they are used only to balance the numerical scales of different terms in the transition cost. We use the same hyperparameters across all experiments.

B.2 Hyperparameters for Hand Refinement

Hand refinement hyperparameters. We run hand refinement on an NVIDIA 4090 GPU. Each action segment is optimized with Adam for 400 steps, using fixed learning rates of 1×10^{-5} for hand scale, 2×10^{-3} for translation, 1×10^{-4} for global orientation, 2×10^{-4} for articulation, and 1×10^{-4} for MANO shape. We set $\lambda_{\text{2d}} = 10$, $\lambda_{\text{sdf}} = 10$, $\lambda_{\text{reg}} = 6$, and $\lambda_{\text{temp}} = 8$. The ray contact weight λ_{match} is 30 for the first 100 SDF warmup steps and 20 afterwards. The SDF contact base weight is 5, the grasping-stage mask multiplier is $\gamma_g = 2$, and the second-order smoothness weight λ_{sm} is 0.5 with stride 1.

C Evaluation Protocol

C.1 Object Pose Tracking Metrics

Object poses are estimated using reconstructed object meshes rather than ground-truth CAD models. For evaluation, we align the predicted poses with the ground truth based on the first frame of each trajectory to account for the mismatch between canonical coordinate frames.

For all object pose estimation metrics, we first compute the metric value within each trajectory and then aggregate the resulting per-trajectory values over the evaluation subset. Let $\mathbf{t}$ and $\hat{\mathbf{t}}$ denote ground-truth and predicted translation, $\mathbf{R}$ and $\hat{\mathbf{R}}$ denote ground-truth and predicted rotation, respectively. Let $\mathcal{M}$ be a set of m sampled object-surface points. Translation error is reported in centimeters and rotation error in degrees.

Translation Error:

$$\|\hat{\mathbf{t}} - \mathbf{t}\|_2 \tag{19}$$

Rotation Error:

$$\arccos\left(\frac{\text{tr}(\hat{\mathbf{R}}^\top \mathbf{R}) - 1}{2}\right) \tag{20}$$

The following ADD and ADD-S definitions normalize surface-point distances by the object diameter. ADD uses corresponding points, whereas ADD-S uses nearest-point matching to account for indistinguishable surface configurations. Table 1 reports the area under the accuracy–threshold curve (AUC) for each distance, rather than the mean distance itself.

ADD:

$$\frac{1}{m}\sum_{\mathbf{x}\in\mathcal{M}}\frac{\|(\hat{\mathbf{R}}\mathbf{x}+\hat{\mathbf{t}})-(\mathbf{R}\mathbf{x}+\mathbf{t})\|_2}{\max_{\mathbf{x}_1,\mathbf{x}_2\in\mathcal{M}}\|\mathbf{x}_1-\mathbf{x}_2\|_2} \tag{21}$$

ADD-S:

$$\frac{1}{m}\sum_{\mathbf{x}_1\in\mathcal{M}}\min_{\mathbf{x}_2\in\mathcal{M}}\frac{\|(\hat{\mathbf{R}}\mathbf{x}_1+\hat{\mathbf{t}})-(\mathbf{R}\mathbf{x}_2+\mathbf{t})\|_2}{\max_{\mathbf{x}_3,\mathbf{x}_4\in\mathcal{M}}\|\mathbf{x}_3-\mathbf{x}_4\|_2} \tag{22}$$

C.2 Hand Refinement Metrics

Following prior hand-object motion reconstruction work, we report Mean Per-Joint Position Error (MPJPE) for hand pose accuracy and Mean Root Position Error (MRPE) for wrist trajectory accuracy after first-frame alignment. We use Interpenetration Volume (IV) to measure hand-object interpenetration, reporting both mean and maximum values, and Jerk Magnitude (JM) to measure temporal smoothness. MPJPE and MRPE are computed on visible-hand frames, while IV and JM are computed on interaction frames. All metrics are lower-better, with units mm for MPJPE/MRPE, cm^3 for IV, and mm/s^3 for JM.

Let $\hat{J}_{t,j}$ and $J_{t,j}$ denote the predicted and ground-truth 3D hand joints at frame t, where $j = 0,\ldots,J-1$ and $J=21$. We use $\hat{r}_t=\hat{J}_{t,0}$ and $r_t=J_{t,0}$ as the predicted and ground-truth wrist/root joints. Pose metrics are computed on visible hand frames $\mathcal{V}$, while penetration and smoothness metrics are computed on interaction frames $\mathcal{I}_{\text{inter}}$.

We report root-aligned MPJPE to evaluate hand pose accuracy independently of global wrist translation:

$$\text{MPJPE}=\frac{1}{|\mathcal{V}|J}\sum_{t\in\mathcal{V}}\sum_{j=0}^{J-1}\left\|(\hat{J}_{t,j}-\hat{r}_t)-(J_{t,j}-r_t)\right\|_2. \tag{23}$$

For wrist/root trajectory accuracy, we use first-frame-aligned MRPE. Let $t_0=\min\mathcal{V}$ and $\tilde{r}_t=\hat{r}_t-\hat{r}_{t_0}+r_{t_0}$. Then

$$\text{MRPE}=\frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}}\|\tilde{r}_t-r_t\|_2. \tag{24}$$

For physical plausibility, we report hand-object interpenetration volume. For each interaction frame, we approximate the intersection volume with a voxel grid $\mathcal{G}_t$ over the overlapping bounding box of the hand mesh $\hat{\mathcal{H}}_t$ and object mesh $\hat{\mathcal{O}}_t$:

$$\text{IV}_t=p^3\sum_{x\in\mathcal{G}_t}\mathbf{1}[x\in\hat{\mathcal{H}}_t]\,\mathbf{1}[x\in\hat{\mathcal{O}}_t], \tag{25}$$

where p is the voxel side length and $\mathbf{1}[\cdot]$ is the occupancy indicator. The resulting volume is reported in cm^3.

For temporal smoothness, we compute jerk magnitude on the predicted wrist/root trajectory. In the definition below, τ denotes physical time in seconds and wrist positions are expressed in millimeters:

$$\text{JM}=\frac{1}{|\mathcal{I}_{\text{inter}}|}\sum_{t\in\mathcal{I}_{\text{inter}}}\left\|\left.\frac{d^3\hat{r}(\tau)}{d\tau^3}\right|_{\tau=\tau_t}\right\|_2. \tag{26}$$

D Extended Reconstruction Evaluation

D.1 Object Reconstruction Quality

We evaluate object reconstruction quality on DexYCB and compare our pipeline with HOLD (Fan et al., 2024) and MagicHOI (Wang et al., 2025). Chamfer distance (CD) measures bidirectional

nearest-surface discrepancy, while surface F-scores combine the precision and recall of matches within a distance threshold. Table 5 reports CD in cm^2 and F@5 and F@10 in percent. Our pipeline obtains the lowest CD and the highest F-scores among the compared methods on the evaluated data.

Table 5: Object reconstruction quality on DexYCB. Lower CD and higher F-scores are better. Best values are highlighted in bold.

Method	CD (cm^2) ↓	F@5 (%) ↑	F@10 (%) ↑
HOLD (Fan et al., 2024)	19.30	33.20	54.94
MagicHOI (Wang et al., 2025)	2.05	45.67	67.14
Ours	**0.57**	**49.35**	**87.54**

Table 6: We ablate the object grounding module. "Valid" denotes the percentage of sequences with a valid prediction and non-zero overlap with the ground-truth mask. Metrics are reported as full/valid, where full counts detection failures and zero-overlap predictions as IoU $= 0$, and valid is computed only over valid sequences.

VLM	Detector	Valid (%) ↑	mIoU ↑	Acc@0.25 (%) ↑	Acc@0.50 (%) ↑	Acc@0.75 (%) ↑
×	×	20.2	0.08/0.38	13.2/65.3	7.3/36.1	2.5/12.4
×	✓	84.5	0.56/0.66	76.0/89.9	70.2/83.1	42.3/**50.1**
✓	×	87.0	0.58/0.67	83.5/96.0	71.1/81.7	38.8/44.6
✓	✓	**95.8**	**0.66/0.69**	**93.8/97.9**	**81.1/84.7**	**44.8**/46.8

D.2 Object Grounding Ablation

We additionally evaluate the object grounding module used before object reconstruction and pose tracking on DexYCB. The full grounding pipeline consists of three steps: a VLM first predicts the name of the manipulated object, an open-vocabulary detector then localizes the object according to the predicted text prompt, and SAM3 finally produces the object mask throughout the video using the detected bounding box as input. To assess the contribution of each component, we compare four variants: directly prompting SAM3 with "object in contact with human hand", using the same generic prompt to drive the detector before SAM3, using the VLM-predicted object text to prompt SAM3, and using the complete VLM-detector-SAM3 pipeline.

All metrics are computed at the sequence level. Specifically, we first compute a local IoU score for each sequence by averaging the per-frame mask IoU within that sequence, and then aggregate the resulting sequence-level scores over the dataset. We report mean IoU and Acc@τ, where Acc@τ denotes the percentage of sequences whose sequence-level IoU is higher than τ. For each metric, we report two values in the form full/valid. The full setting includes all sequences and assigns IoU = 0 to detection failures and zero-overlap predictions. The valid setting is computed only over sequences for which the method produces a valid prediction with non-zero overlap with the ground-truth mask. We also report Valid, defined as the percentage of such valid sequences over the full evaluation set.

As shown in Table 6, directly prompting SAM3 with a generic interaction description performs poorly, achieving only 20.2% valid sequences and an overall mIoU of 0.08. This indicates that mask propagation alone is insufficient when the target object is not explicitly localized. Introducing either detector-based localization or VLM-based object naming substantially improves grounding performance: the detector-only variant increases the valid rate to 84.5%, while the VLM-only variant reaches 87.0%. The full pipeline achieves the best overall performance, with 95.8% valid sequences, an overall mIoU of 0.66, and the highest full-setting Acc@0.25, Acc@0.50, and Acc@0.75. These gains show that VLM-based semantic object identification and detector-based spatial localization are complementary: the VLM provides a more precise object-level prompt, while the detector supplies an explicit spatial prior for SAM3. Although the detector-only variant obtains a slightly higher valid-setting Acc@0.75, its lower valid rate and weaker full-setting results suggest that it is less reliable as an end-to-end grounding module. Overall, the complete design improves both the coverage of successful grounding and the quality of the resulting object masks.

D.3 Hand Refinement Ablation

Effectiveness of optimization components. We conduct an ablation study on DexYCB to inspect the contribution of the main optimization components. Table 7 reports the full objective and variants with individual losses removed.

The full objective reflects the combined effect of all optimization constraints. Removing the 2D loss reduces IV-mean from 1.07 to 0.27, but increases MRPE from 38.49 to 113.17 and JM from 2.49 to 9.93. This highlights the role of image alignment in constraining hand motion and shows that lower penetration alone does not imply more accurate interaction reconstruction. Removing the ray contact loss weakens the direct hand-object coupling, making the trajectory less constrained by interaction evidence. Removing the temporal ray contact loss increases JM from 2.49 to 3.19, indicating that this constraint improves wrist trajectory smoothness. The full objective balances geometric accuracy, contact quality, and temporal stability.

Table 7: Ablation of the optimization losses on DexYCB. We report MPJPE, MRPE, and physics-based measures. The first row uses the full objective, while each following row removes one loss term to assess the role of each constraint.

$\mathcal{L}_{\text{2d}}$	$\mathcal{L}_{\text{match}}$	$\mathcal{L}_{\text{temp}}$	MPJPE ↓	MRPE ↓	IV-mean ↓	IV-max ↓	JM ↓
✓	✓	✓	**7.58**	**38.49**	1.07	3.25	**2.49**
×	✓	✓	8.38	113.17	**0.27**	**1.55**	9.93
✓	×	✓	7.64	48.70	0.53	2.27	2.53
✓	✓	×	**7.58**	38.62	1.07	3.12	3.19

Additional Baseline. To further evaluate the effectiveness of our hand reconstruction refinement, we include MagicHOI as an additional baseline as shown in Table 8. Since our goal is to compare different hand refinement strategies under the same reconstruction pipeline, we adopt the hand optimization module from MagicHOI while keeping all other components identical to our method, including object reconstruction, object pose estimation, and initialization procedures. This setup isolates the effect of the hand refinement stage and enables a more controlled comparison between the two optimization strategies. As shown in Table 8, MagicHOI performs worse than our refinement across all reported metrics and even underperforms the WiLoR initialization on several metrics.

Table 8: We compare additional baseline MagicHOI with WiLoR initialization and our refined hand trajectories on HO3D-v3. Best values are highlighted in bold.

Dataset	**Method**	MPJPE ↓	MRPE ↓	IV-mean ↓	IV-max ↓	JM ↓
HO3D-v3	MagicHOI	5.90	22.64	13.24	48.80	1.25
	WiLoR	**4.22**	24.22	20.11	64.82	0.86
	Ours	4.40	**16.99**	**3.42**	**22.40**	**0.34**

E Qualitative Results

E.1 Hand-Object Reconstruction on HO3D-v3

Figure 4 shows qualitative results on HO3D-v3. Our method consistently improves interaction quality over both the initialization and MagicHOI for objects with varying geometric complexity, ranging from simple box-shaped objects to challenging drill-shaped objects. In the drill example, the initialization and MagicHOI do not recover a valid grasp, whereas our method successfully reconstructs the hand as grasping the drill. Despite slight residual interpenetration, the reconstructed interaction remains substantially more plausible and stable. The proposed refinement effectively alleviates hand-object interpenetration and false-contact artifacts, resulting in more physically consistent reconstructions. As highlighted by the green circles, our method occasionally produces interaction configurations that appear visually more consistent with the image observations than the provided annotations.



Figure 4: Qualitative results of hand-object reconstruction on HO3D-v3. The reconstruction result is visualized in both the front and side views.

F Real-World Robot Experiment Details



Figure 5: Hardware setup of real-world robot experiments.

Figure 5 shows the Franka arm and xHand used for the real-robot evaluation. The current comparisons cover Place Vase, Open Box, and Insert Toast (Figure 3). Human and robot demonstrations are recorded with the same fixed side-view camera system.

Action format. The action predicted by the policy consists of two components. The first component represents the pose of the robot end-effector, including a 3D translation and a 3D rotation parameterized in axis-angle form. The second component corresponds to the joint angles of the dexterous hand. In our implementation, we use xHand, which has 12 degrees of freedom. Therefore, the complete action vector has 18 dimensions. The policy predicts relative action offsets rather than absolute target values.

For WiLoR, MagicHOI, and Hand2Robot, we apply the same xHand retargeting, coordinate transformation, and relative-action encoding procedures to the recovered human hand motion before policy post-training.

For Do As I Do (Paliwal et al., 2026), we retain its dynamics-based post-optimization and adapt the optimized Sharpa motion to xHand using the same retargeting method as Hand2Robot.

Training data and sampling. We post-train $\pi_{0.5}$ for 30,000 steps using the per-task demonstration budgets in Tables 3 and 4. The action-generation comparison uses the same 20 robot demonstrations and 100 human source videos per task for WiLoR, MagicHOI, and Hand2Robot. The budget study evaluates 20, 40, and 60 robot demonstrations, together with mixed-data conditions using 20 robot demonstrations and either 50 or 100 human demonstrations. At 20R+50H, Do As I Do and Hand2Robot process the same 50 source human videos per task. The human-only condition uses 100H.

For all mixed-data conditions, human and robot training samples are drawn at a 1:1 ratio, regardless of their relative demonstration counts. Robot-only and human-only conditions sample from their respective data sources.

Object placements and evaluation. Objects are placed on a tabletop workspace measuring 40 cm $\times$ 70 cm. In the action-generation comparison in Table 3, human and robot demonstrations cover the same spatial region. Each trained policy is evaluated in 20 trials per task, with initial object positions randomly sampled from the training distribution. We compute task success rate as the number of successful trials divided by 20, and give equal weight to the three tasks when reporting the mean.

Task success. Success is judged by task completion: placing the vase on the plate in Place Vase, opening the box in Open Box, and inserting the toast into the toaster in Insert Toast. The final row of Figure 3 illustrates successful completion of the three tasks.