arXiv:2610.41336v1 [cs.RO] 30 Oct 2026

HandMimicX: Learning Human-Like Dexterous Manipulation via an Object-Centric Policy from Egocentric Videos

Zihan Gao[1,2,3,*] Yufei Liu[2,4,*] Yuhong Zhang[1,2,‡] Zhen Yu[2] Xinhao Huang[1,2]
Zhinan You[2] Kai W. Ho[5] Zipeng Ye[3,†] Peng Jiao[1,2] Haoqian Wang[1,†]

[1]Tsinghua University [2]TsingTec [3]MeiTuan
[4]The Chinese University of Hong Kong, Shenzhen [5]NVIDIA



Fig. 1: ***HandMimicX***: We present an object-centric framework for learning dexterous manipulation skills from a single egocentric demonstration, leveraging simulation-based reinforcement learning and sim-to-real policy transfer. The framework consists of three stages. First, we reconstruct the demonstrated hand–object interaction from the input video, recovering the hand–object motion trajectory, object geometry, and task-relevant interaction priors. Second, using the reconstructed object mesh and interaction priors, we train an object-centric tracking policy to manipulate the object to spatially randomized target object poses, denoted as RT_n. By optimizing the policy across diverse target poses, the robot acquires spatial generalization beyond the original human demonstration. Finally, the learned policy is transferred to the real robot and executed using observations from an RGB-D camera, enabling dexterous manipulation in the real-world environment.

***Abstract*— Learning dexterous manipulation from human videos remains an open challenge, particularly in the context of transferring a single demonstration into a spatially generalizable robotic skill and maintaining physically plausible human-like interactions. Unlike prior works that emphasize hand-motion imitation, we introduce an object-centric framework for acquiring dexterous skills from egocentric human RGB-D videos. Within our framework, we formulate the reinforcement learning (RL) objective in terms of object motion to decouple the underlying skill from the specific hand trajectory and impose constraints on hand–object interaction states to promote more anatomically natural hand configurations. Extensive experiments in simulation and on real robots demonstrate that our method enables efficient learning from human videos and generalizable skill transfer across novel spatial configurations. Project page: `https://anonymous.4open.science/w/handmimicx/index.html`.**

*Co-first authors. ‡Project lead. †Corresponding author.
Email: {gaozihanthu, liuyufei200003}@gmail.com

I. Introduction

Recent advances in embodied intelligence have demonstrated the potential of learning robot manipulation policies from human videos [1], [2], [3], [4]. However, directly transferring human demonstrations to robots remains challenging due to the gap between human motion and robot embodiments, as well as the discrepancy between visual observations and physical execution [5], [6], [7].

Reinforcement learning (RL) provides a powerful mechanism to bridge this gap by optimizing interaction behaviors under physical constraints in simulation. Under this paradigm, human demonstrations can be transformed into physically feasible robot manipulation policies. However, acquiring a robust and generalizable skill from a single human demonstration remains challenging. Existing RL approaches largely rely on hand motion as the primary supervision signal, learning residual actions to refine the

reference motion [5], [8], [9], [10]. While this paradigm enables precise motion imitation, it relies heavily on accurately reconstructed hand–object trajectories, making performance sensitive to even minor misalignments and favoring kinematic reproduction over the acquisition of transferable skills. By contrast, object-centric policies [11], [12] prioritize desired object states, shifting the learning objective from replicating a demonstration to capturing what the task is meant to accomplish. Nevertheless, their disproportionate emphasis on object-level task success often leads to unrealistic hand–object interactions, including physically implausible contacts and unnatural finger configurations.

In light of these challenges, we propose **HandMimicX**, a novel object-centric reinforcement learning framework for acquiring an underlying manipulation skill from a single egocentric human video. The "X" indicates moving beyond trajectory mimicry toward generalizable skill acquisition. To bridge egocentric human videos and dexterous control, we jointly reconstruct hand-object interactions (HOIs). However, camera motion degrades reconstruction accuracy, particularly in terms of hand–object alignment. We therefore refine the reconstructed hand translation using depth to improve spatial consistency. To transcend the trajectory constraints inherited from the human demonstration video, we condense the full-sequence reconstructions into object geometry and interaction priors to retain the most critical task-relevant elements. Using the reconstructed geometry, we train an RL policy to manipulate the object to arbitrary 6-DoF target poses, while the interaction prior regularizes the resulting hand–object interactions toward physically plausible configurations. For real-world deployment, we use DAgger [13] to distill the policy and reduce its reliance on privileged information. Ultimately, our framework transforms a single human demonstration video into a transferable dexterous skill that generalizes beyond the demonstrated spatial configuration. **In summary, our contributions are as follows:**

- We propose **HandMimicX**, a novel object-centric RL framework that learns a generalizable dexterous manipulation skill from a single human video.
- We build a joint 4D HOI reconstruction pipeline that recovers spatially consistent hand–object motion from dynamic egocentric views.
- We enhance object-centric RL with an extracted interaction prior, regularizing policy learning toward physically plausible and anthropomorphic dexterous manipulation.

II. Related Work

A. Human–Object Interaction Reconstruction

Recovering human–object interactions (HOIs) from visual observations provides the geometric interface between human videos and downstream robot learning [14]. HOLD [15] leverages a compositional articulated implicit model to jointly reconstruct articulated hands and objects using compositional SDF and contact constraints. HO-NeRF [16] separately learns shape and appearance prior knowledge of hands and objects with a neural representation and proposes a pose estimation system for hand-object interaction from sparse views. These methods, however, rely on either scene-specific optimization or pre-built hand models, limiting their scalability to in-the-wild videos. Another line of work, exemplified by HaMeR [17], scales up a ViT-based backbone for 3D hand reconstruction, yielding accurate and robust hand mesh recovery, while HaWoR [18] and Dyn-HaMR [19] couple camera tracking with hand-motion estimation for global 4D hand recovery from unconstrained videos. Recent studies [14], [20], [21] have developed more generalizable integrated frameworks for reconstructing overall hand–object motion. Nevertheless, these approaches remain susceptible to cumulative errors, particularly in hand–object alignment, compromising downstream manipulation.

B. Learning Robot Manipulation from Human Videos

Exploiting human videos to train robotic models has been actively studied in recent years. Recently, [1], [2], [22], [23] have utilized human video for pretraining robotic manipulation models; for instance, EgoScale [1] uncovers a clear log-linear relationship between data scale and hand-action prediction loss and demonstrates that this loss strongly correlates with real-robot dexterous manipulation performance. By leveraging large-scale human video corpora [24], [25], [26], [27], [28], these works envision a future in which robots acquire diverse skills from abundant human demonstrations. In parallel, advances in simulation [29], [30] have established real-to-sim-to-real as an alternative paradigm for robot learning from videos [3], [4], [31], [32], in which policies are optimized via RL to reproduce motions demonstrated in human videos. Our proposed **HandMimicX**, instead, emphasizes learning a transferable skill via RL from a single human demonstration video, moving beyond mere imitation of the demonstrated trajectories.

C. Object-Centric Reinforcement Learning

Unlike residual RL [8], [32], [33], which remains anchored to a reference trajectory, object-centric RL [34], [35] directly optimizes task-relevant object-state transitions. Dex4D [11] learns a generalizable task-agnostic grasping policy, while X-Sim [36] trains a reinforcement learning (RL) policy in photorealistic simulations reconstructed from RGB-D human videos. To simplify RL for robotic hand control, Human2Sim2Robot [37] combines geometric fabric control with a PCA-based action space, but this formulation can lead to degenerate hand configurations. These prior methods either target grippers or neglect the naturalness and anthropomorphism of hand configurations during grasping. By contrast, we transfer physically grounded hand–object interaction patterns from human videos to promote feasible dexterous manipulation.

III. Method

A. Overview

An overview of our framework is shown in Fig. 2. Given a single egocentric RGB-D video, we reconstruct a depth-refined HOI trajectory to extract the metric-scale object mesh

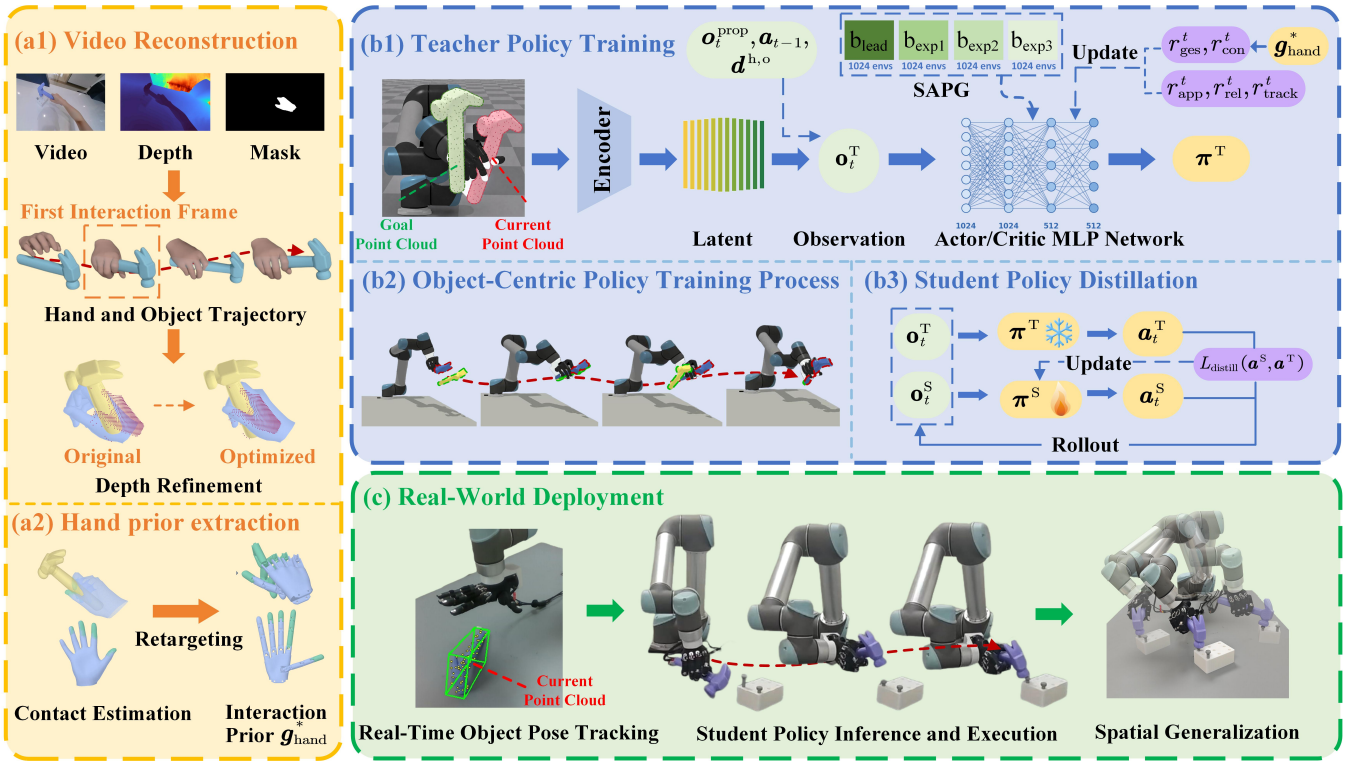


Fig. 2: Overview of **HandMimicX**. (a1) Given an RGB-D egocentric video, we reconstruct the manipulated object and hand motion, and then align both reconstructions with the depth observations. (a2) We extract the hand-object interaction state at the first contact frame as the interaction prior g^*_{hand} for policy learning. (b1) We train an object-centric teacher policy π^{T} using PPO with SAPG. A correspondence-aware encoder maps the current and goal object point clouds to a geometric latent representation, which is concatenated with o_t^{prop}, a_{t-1} and $d^{\text{h,o}}$ to form the observation o_t^{T} and then fed into the actor and critic MLPs. (b2) During training, the teacher π^{T} learns to manipulate the object and track arbitrary goal poses. (b3) We use DAgger to distill the teacher π^{T} into a student π^{S} under partial observations o_t^{S}. Finally, (c) we deploy the distilled student policy on the real robot to execute diverse trajectories, extending beyond the original video demonstration.

and the interaction prior. Next, we train an object-centric policy under the constraints of the prior in simulation. For deployment, we use DAgger [13] to distill the privileged policy trained in simulation into a student policy conditioned on partial observations available in the real world.

B. Egocentric HOI Reconstruction

1) Problem Setup: Given a dynamic egocentric video sequence $I_{t_{t=1}}^{T}$, we formulate the training objective based on the following physical state representations: the geometry of the manipulated object and interaction priors capturing hand configuration, hand–object relative pose, and contact.

2) Joint HOI Reconstruction with Depth Refinement: We aim to recover spatially and temporally consistent hand–object interactions from dynamic egocentric observations. Specifically, we use SAM3D [38] to recover the metric-scale object mesh, HaWoR [18] to estimate hand poses in the world coordinate frame, and HACO [39] to detect hand–object contact and identify the corresponding contact regions, as shown in Fig. 3. The interaction prior is then defined by the hand pose, relative hand–object position, and region-level contact at contact onset. We additionally reconstruct the demonstrated object trajectory with FoundationPose [40] throughout the video. Although this trajectory is not directly used for policy training, it provides a reference for real-world execution and serves as a basis for generating additional object motion trajectories.

However, a key challenge is that independently reconstructing the hand and object under continuous ego-camera motion can introduce scale drift and spatial misalignment, leading to inconsistent interaction trajectories. To mitigate the inconsistency between reconstructed hand and object trajectories, we further refine the global hand translation using metric depth observations. For each interaction frame t, we back-project valid depth pixels within the hand mask into a world-space point cloud and randomly sample 10 points to form a sparse set of depth anchors $\mathcal{P}_t$. We then identify the MANO mesh vertices visible to the camera and obtain their



Fig. 3: **Visualization of contact and hand pose retargeting.** We use HACO to estimate contact regions on the MANO hand mesh and transfer them to the robotic hand for contact retargeting. Hand pose retargeting is performed using [41].

world-space coordinates, denoted by $\mathcal{V}_t$. The global hand translation is refined by aligning the visible hand surface with the depth observations:

$$\Delta \mathbf{t}_t^* = \arg\min_{\Delta \mathbf{t}} \sum_{\mathbf{v} \in \mathcal{V}_t} \rho\left(\min_{\mathbf{p} \in \mathcal{P}_t} \|\mathbf{v} + \Delta \mathbf{t} - \mathbf{p}\|_2^2 \right). \quad (1)$$

We adopt the Huber loss $\rho(\cdot)$ to reduce sensitivity to depth outliers. The refined hand translation is $\mathbf{t}_t = \mathbf{t}_t^0 + \Delta \mathbf{t}_t^*$. By anchoring the hand reconstruction to metric depth observations, this refinement alleviates scale drift and hand–object misalignment caused by ego-camera motion, substantially reducing interpenetration and improving the spatial and temporal consistency of the recovered interaction.

C. Object-Centric Policy for Learning from Videos

1) Object-Centric Policy Design: Our policy $\boldsymbol{\pi}$ employs an actor–critic architecture to jointly control the 6-DoF UR5 arm and the 20-DoF Wuji hand. At each time step t, the actor receives an observation $\boldsymbol{o}_t$ consisting of the current and target object point clouds, robot joint positions and velocities $\boldsymbol{o}_t^{\text{prop}}$, the previous action $\boldsymbol{a}_{t-1}$, and distances $\boldsymbol{d}^{\text{h,o}}$ from the fingertips to their nearest object points. We employ a correspondence-aware encoder to capture geometric relationships between the current and target object point clouds while preserving pointwise correspondences, yielding a compact and expressive latent representation of their relative pose. The remaining observations are concatenated and passed through an MLP encoder to obtain a proprioceptive embedding. All embeddings are jointly fed into the actor to produce a 26-dimensional action vector. The predicted action is executed in the simulation environment to interact with the object and receive the corresponding reward. Our critic estimates the expected cumulative future return from the current state under the actor policy, providing a value signal to guide policy optimization. Throughout the process, the interaction prior shapes the reward by measuring the deviation from the reference hand pose, encouraging the policy to maintain physically plausible hand configurations. Additional privileged information, including contact and collision signals, is used to assess policy performance.

a) Point Correspondence Learning: Motivated by [42], we use point correspondences to capture the geometric relationship between the current and goal object states. Let $\bar{\mathbf{X}} = \{\bar{\mathbf{x}}_i\}_{i=1}^N$ denote a set of $N = 128$ points sampled from the object mesh. A shared PointNet [43] encoder ϕ_θ extracts pointwise embeddings from the current and goal point clouds: $\mathbf{H}_t^s = \phi_\theta(\mathbf{X}_t^s)$, where $s \in \{c, g\}$, with c and g denoting the current and goal object point clouds, respectively. To capture the correspondence between the current and goal object states, we apply bidirectional cross-attention:

$$\begin{aligned} \widetilde{\mathbf{H}}_t^c &= g_\theta(\mathbf{H}_t^c, \mathbf{H}_t^g) + \mathbf{H}_t^c, \\ \widetilde{\mathbf{H}}_t^g &= g_\theta(\mathbf{H}_t^g, \mathbf{H}_t^c) + \mathbf{H}_t^g. \end{aligned} \quad (2)$$

Here, $\mathbf{H}_t^c$ and $\mathbf{H}_t^g$ denote the pointwise features of the current and goal object point clouds, respectively, and g_θ denotes the transformer-based cross-attention module. The resulting features $\widetilde{\mathbf{H}}_t^c$ and $\widetilde{\mathbf{H}}_t^g$ mutually encode the two geometries, capturing their correspondence and producing interaction-aware representations for fusion.

b) Reward Functions: The total reward is designed to encourage the robotic hand to manipulate the object toward the target poses while promoting plausible interactions:

$$\begin{aligned} r_t = w_{\text{app}} r_{\text{app}} + w_{\text{con}} r_{\text{con}} \\ + w_{\text{ges}} r_{\text{ges}} + w_{\text{rel}} r_{\text{rel}} + w_{\text{track}} r_{\text{track}}. \end{aligned} \quad (3)$$

The approach reward r_{app} encourages the robotic hand to approach the object by reducing the finger–object distance $\boldsymbol{d}^{\text{h,o}}$. The contact reward r_{con} encourages the target regions of the robotic hand to make contact with the object.

To identify the target contact regions, we first use HACO to estimate contacts on the human hand and then retarget them to the robotic hand, as illustrated in Fig. 3. Specifically, HACO predicts a binary contact label $h_t(\boldsymbol{v}) \in \{0, 1\}$ for each MANO vertex $\boldsymbol{v}$. We aggregate these vertex-level predictions into region-level contact estimates using a predefined anatomical segmentation. These contact estimates are then transferred to the corresponding semantic regions on the Wuji hand, each associated with a set of candidate robot links. The resulting set of target contact regions, denoted by $\mathcal{C}$, is used to define the contact reward:

$$r_{\text{con}} = \sum_{i \in \mathcal{C}} I\left(d_i^{\text{h,o}} < \epsilon \right), \quad (4)$$

where $d_i^{\text{h,o}}$ denotes the distance between the i-th robotic hand region and the object, ϵ is the contact threshold, and $I(\cdot)$ is the indicator function. The relative-pose reward r_{rel} preserves the demonstrated hand–object configuration by penalizing deviations in both relative translation $\mathbf{P}_{\text{rel}}^t - \hat{\mathbf{P}}_{\text{rel}}^t$ and rotation $\mathbf{R}_{\text{rel}}^t \ominus \hat{\mathbf{R}}_{\text{rel}}^t$, where $\mathbf{P}_{\text{rel}}^t$ and $\mathbf{R}_{\text{rel}}^t$ denote the relative position and orientation, respectively; hats indicate demonstration references, and $\ominus$ denotes the rotation difference expressed as an axis–angle vector.

The gesture reward r_{ges} promotes alignment of the robot's finger configuration with the demonstrated gesture. We retarget the demonstrated hand pose to the Wuji hand using [41], as shown in Fig. 3, and then represent the gesture as a set of key vectors extending from the palm center to the finger joints. For finger f and joint j, the vector is defined as

$$\mathbf{v}_{f,j}^t = (\mathbf{R}_t^H)^\top \left(\mathbf{p}_{f,j}^t - \mathbf{p}_t^H \right), \quad (5)$$

where $(\mathbf{p}_t^H, \mathbf{R}_t^H)$ denotes the current palm pose. Given the corresponding reference vector $\mathbf{v}_{f,j}^*$, we formulate the gesture reward as

$$r_{\text{ges}}^t = \exp\left(-\beta \sum_{f,j} \left\| \mathbf{v}_{f,j}^t - \mathbf{v}_{f,j}^* \right\|_2 \right), \quad (6)$$

where β controls the sensitivity to gesture deviation. This reward assigns higher values to hand configurations that more closely match the reference interaction gesture. Finally, r_{track} encourages tracking of target positions by awarding a bonus upon reaching each target.



Fig. 4: Visualization of sampled objects for policy pretraining. The collection spans diverse categories, including everyday objects and tools, with varied geometries.

c) Training Strategy: The high-dimensional action space of the robotic hand makes direct policy optimization challenging. We therefore adopt a two-stage training scheme consisting of pretraining and finetuning. The pretraining stage equips the policy with general manipulation capabilities and provides a favorable initialization for downstream learning by training on over 3200 objects [25], [44], as shown in Fig. 4. Each object is associated with hundreds of reference hand poses. We then finetune the policy on each task to acquire the desired skills, guided by the object and hand–object interaction priors extracted from the corresponding HOI video. Notably, pretraining and finetuning share the same policy architecture and reward formulation. We pretrain the policy for 30K iterations and finetune for 10K per task across 4,096 parallel environments. SAPG [45] is used to optimize both stages, with its four blocks employing distinct exploration distributions to collect diverse rollouts and broaden coverage of the action space.

2) DAgger Distillation Training: Given the optimized teacher policies after finetuning, we distill them into deployable student policies via DAgger [13], enabling real-world execution under partial observations. The student policy adopts the same actor architecture as the teacher. It excludes finger–object distances and contact information and is trained to match the teacher's actions by minimizing the objective:

$$\mathcal{L}_{\text{distill}} = \left\| \pi^{\text{S}}\left(o_t^{\text{partial}}\right) - \pi^{\text{T}}\left(o_t^{\text{priv}}\right) \right\|_2^2, \tag{7}$$

where o_t^{priv} and o_t^{partial} denote the privileged and real-world-accessible observations, respectively.

IV. Experiments

A. Experimental Setup

We benchmark our method against two object-centric approaches, SimToolReal and Dex4D, across eight manipulation tasks. For each baseline, we follow its original training protocol and introduce an additional finetuning stage for task-specific adaptation. The tasks span diverse hand–object interactions, including tool use, object transport, placement, and stacking. We briefly describe each task below.

- **Nail Hammering**: Grasp a hammer to strike a nail.
- **Umbrella Hanging**: Hang an umbrella on a target hook.
- **Debris Sweeping**: Sweep debris into a dustpan.
- **Pepper Lifting**: Stably lift a pepper from the tabletop.
- **Eggplant Placement**: Place an eggplant in a basket.
- **Bowl Stacking**: Stack one bowl on top of another.
- **Cup Relocation**: Place a cup on the wooden rack.
- **Upright Sprayer**: Reorient a spray bottle upright.

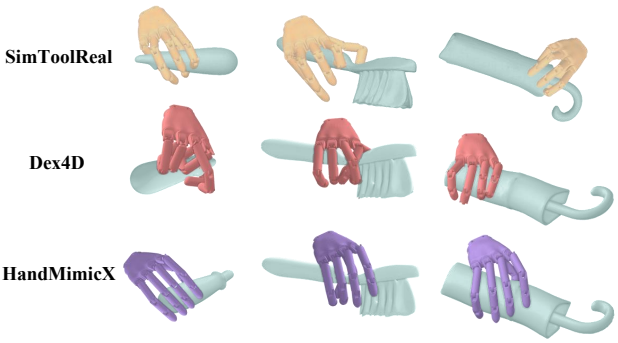


Fig. 5: Visualization of hand–object interaction poses from SimToolReal, Dex4D, and HandMimicX. Our HandMimicX produces the most anatomically natural hand poses.

B. Experiments in Simulation

We compare our method with the baselines in simulation by evaluating their ability to move objects through a sequence of randomly sampled waypoints, following the evaluation protocol of Dex4D and SimToolReal. The target advances to the next waypoint once the current waypoint is reached, and a trial is considered successful if all 5 waypoints are reached. Success rates are computed across 4,096 environments. As reported in Tab. I, our method achieves substantially higher success rates than the baselines, with an average success rate of 85.4%.

a) Visualization of Hand-Object Interaction Poses: In addition to improving object manipulation success rates, our interaction formulation plays a more fundamental role in maintaining natural, human-like hand–object configurations. As illustrated in Fig. 5, our method suppresses implausible hand postures, yielding more realistic finger articulation and appropriate contact with the object.

b) Ablations: We conduct comprehensive ablation studies to systematically assess the contributions of point correspondence and the interaction prior. Specifically, we separately remove the feature fusion module from the correspondence-aware encoder and the interaction prior regularization to form two ablated variants, while keeping all other training settings unchanged. These ablations isolate the role of point correspondence in encoding the relative geometry between the current and target object configurations, as well as the effect of the interaction prior on maintaining plausible and desired hand poses. Tab. II shows that both components improve performance individually and yield further gains when combined. To further assess the contribution of pretraining, we conduct an additional ablation study comparing policies trained with and without pretraining. The results show that pretraining increases the average success rate by 12.1%, suggesting that it provides a more effective initialization for subsequent policy optimization.

TABLE I: Success rates (%) across eight manipulation tasks in simulation, with averages in the final column.

Policy	Nail Hammering	Umbrella Hanging	Debris Sweeping	Upright Sprayer	Cup Relocation	Bowl Stacking	Eggplant Placement	Pepper Lifting	Avg.
SimToolReal	39.2	48.2	42.3	64.6	64.4	69.0	77.1	53.6	57.3
Dex4D	57.8	42.7	75.3	78.4	69.9	58.0	89.6	62.5	66.8
HandMimicX	**90.5**	**83.9**	**85.7**	**81.9**	**84.9**	**75.2**	**92.9**	**87.8**	**85.4**

TABLE II: Ablation study evaluating the contributions of the interaction-pose prior, point-correspondence module, and policy pretraining to the overall manipulation performance of our framework. Results are reported as task success rates (%) across different ablation settings, with averages in the final column.

Policy	Nail Hammering	Umbrella Hanging	Debris Sweeping	Upright Sprayer	Cup Relocation	Bowl Stacking	Eggplant Placement	Pepper Lifting	Avg.
w/o Pose Prior	78.3	64.1	79.4	64.5	69.7	76.6	92.5	75.9	75.1
w/o Correspondence	85.8	81.7	82.3	72.1	78.5	64.4	87.4	74.2	78.3
w/o Pretraining	79.7	71.6	62.2	71.3	75.0	61.8	90.6	74.1	73.3
HandMimicX	**90.5**	**83.9**	**85.7**	**81.9**	**84.9**	**75.2**	**92.9**	**87.8**	**85.4**



Fig. 6: Visualization of real-world execution with HandMimicX. Object trajectories extracted from human demonstration videos serve as targets for robot execution (please refer to the accompanying video for further visualizations).

TABLE III: Skill transfer evaluation across eight real-world manipulation tasks. Each cell reports successful trials / total trials. The final column shows the average success rate.

Nail Hammering	Umbrella Hanging	Debris Sweeping	Upright Sprayer	Cup Relocation	Bowl Stacking	Eggplant Placement	Pepper Lifting	Avg. (%)
8/10	7/10	7/10	6/10	7/10	6/10	8/10	8/10	71.3

C. Real-World Deployment

As shown in Fig. 7, our system comprises a 6-DoF UR5 robotic arm equipped with a 20-DoF Wuji hand. A single side-mounted Intel RealSense D435 camera provides visual observations for object tracking. We calibrate the camera extrinsics with respect to the robot base frame. During inference, FoundationPose estimates the object's 6-DoF pose. Points sampled from the object mesh are transformed into the robot base frame using calibrated camera extrinsics to obtain the current object point clouds, synchronized with robot proprioception. The goal point cloud is predefined according to the manipulation objective. The student policy is conditioned on these inputs to predict the next action.

Our real-world evaluation examines whether the proposed object-centric policy captures the manipulation skills underlying human video demonstrations and generalizes their execution across spatially varying trajectories. For each dex-

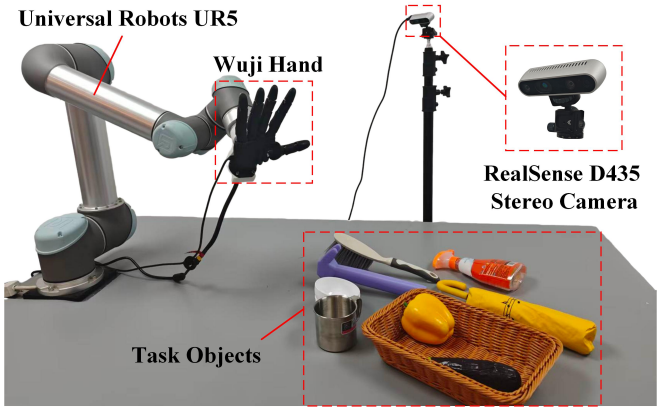


Fig. 7: Real-world hardware setup and objects used in our experiments. The system comprises a 6-DoF UR5 robotic arm equipped with a 20-DoF Wuji hand and a single side-mounted Intel RealSense D435 stereo camera.

terous manipulation task, we conduct 10 trials, each of which requires the policy to execute a distinct object trajectory derived from the demonstrated motion. Specifically, we apply affine transformations to the demonstrated trajectory to vary its intermediate waypoints and target object pose. Fig. 6 shows real-world executions of the object trajectories demonstrated in the videos for all eight tasks, while Fig. 8 illustrates spatial generalization through the execution of diverse trajectories. The quantitative results in Tab. III substantiate the

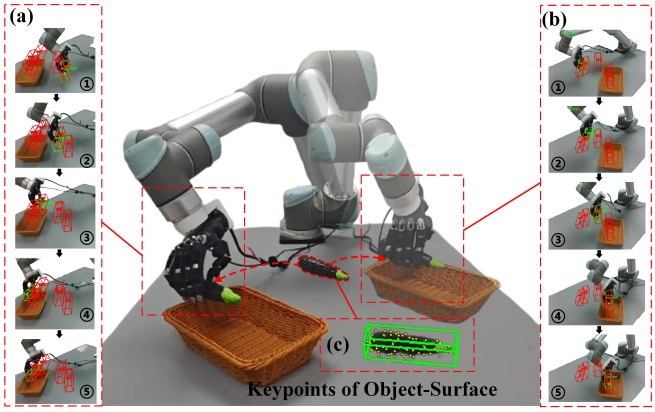


Fig. 8: Skill transfer from a single object trajectory extracted from egocentric video. Panels (a) and (b) show executions from different initial object positions and orientations. Panel (c) visualizes the 128 canonical surface keypoints used to represent the object state for the actor.

effective transfer of the learned generalizable manipulation skills to the physical system.

V. Conclusion

We presented HandMimicX, an object-centric reinforcement learning framework for acquiring a dexterous manipulation skill from a single egocentric human demonstration. Our framework integrates depth-refined hand–object reconstruction, correspondence-aware object representations, and interaction priors that regularize contact formation and finger configurations. By coupling object-level goals with demonstrated interaction structure, HandMimicX enables the robot to adapt its manipulation behavior to spatial configurations beyond the original demonstration. Experiments across eight tasks show improved success rates over object-centric baselines, while ablations support the contributions of interaction priors, geometric correspondence, and pretraining. Real-world deployment also demonstrates the feasibility of transferring these learned behaviors to physical manipulation.

VI. Limitations

Despite its strong performance, our framework has several limitations. It relies on reconstructed object geometry, estimated hand poses, and contact predictions, whose errors can propagate into policy learning. Moreover, our contact representation is limited to discrete region-level states, without modeling finer-grained physical contact forces. Our evaluation of the framework across different embodiments and on bimanual manipulation tasks also remains limited in scope. Future work will investigate more robust human–object interaction perception pipelines, incorporate force-aware contact representations, and extend the framework to more embodiments and manipulation tasks.

References

[1] R. Zheng, D. Niu, Y. Xie, J. Wang, M. Xu, Y. Jiang, F. Castañeda, F. Hu, Y. L. Tan, L. Fu *et al.*, "Egoscale: Scaling dexterous manipulation with diverse egocentric human data," *arXiv preprint arXiv:2602.16710*, 2026.

[2] H. Luo, Y. Feng, W. Zhang, S. Zheng, Y. Wang, H. Yuan, J. Liu, C. Xu, Q. Jin, and Z. Lu, "Being-h0: vision-language-action pretraining from large-scale human videos," *arXiv preprint arXiv:2507.15597*, 2025.

[3] P. Mandikal and K. Grauman, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *Conference on Robot Learning*. PMLR, 2022, pp. 651–661.

[4] Y. Qin, Y.-H. Wu, S. Liu, H. Jiang, R. Yang, Y. Fu, and X. Wang, "Dexmv: Imitation learning for dexterous manipulation from human videos," in *European Conference on Computer Vision*. Springer, 2022, pp. 570–587.

[5] Z. Yuan, T. Wei, L. Gu, P. Hua, T. Liang, Y. Chen, and H. Xu, "Hermes: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation," *ArXiv*, vol. abs/2508.20085, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:280919277

[6] G. Zhang, Q. Xu, H. Zhang, J. Ma, L. He, Y. Bao, Z. Ping, Z. Yuan, C. Lu, C. Yuan, T. Liang, X. Tian, M. Shao, F. Zhang, M. Ding, Y. Gao, H. Zhao, H. Zhao, and H. Xu, "Unidex: A robot foundation suite for universal dexterous hand control from egocentric human videos," *ArXiv*, vol. abs/2603.22264, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:286766691

[7] Y. Han, J. Qiu, L. Bai, Z. Xiao, Z. Zeng, Y. Liu, Z. Yang, S. Jain, W. Ma, J. Fu, Y. Zheng, M. Natarajan, M. Z. Irshad, K. Shaw, M. C. Gombolay, Z. Kira, and H. chaandar Ravichandar, "Video2sim2real: Full-stack autonomous dexterous skill acquisition from a single human video," *ArXiv*, vol. abs/2606.08828, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:289099227

[8] K. Li, P. Li, T. Liu, Y. Li, and S. Huang, "Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning," *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6991–7003, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277435197

[9] X. Liu, J. Adalibieke, Q. Han, Y. Qin, and L. Yi, "Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references," *ArXiv*, vol. abs/2502.09614, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:276317383

[10] Y. Feng, N. Leung, J. Wang, L.-J. Yang, H. Qi, and P. Culbertson, "A minimalist retargeting-guided reinforcement learning recipe for dexterous manipulation," *ArXiv*, vol. abs/2607.11874, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:290143227

[11] Y. Kuang, S. Park, K. Fragkiadaki, and S. Tulsiani, "Dex4d: Task-agnostic point track policy for sim-to-real dexterous manipulation," *ArXiv*, vol. abs/2602.15828, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:285659883

[12] K. Kedia, T. G. W. Lum, J. Bohg, and C. K. Liu, "Simtoolreal: An object-centric policy for zero-shot dexterous tool manipulation," *ArXiv*, vol. abs/2602.16863, 2026. [Online]. Available: https://api.semanticscholar.org/CorpusID:285787326

[13] S. Ross, G. Gordon, and D. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, vol. 15. PMLR, 2011, pp. 627–635.

[14] Y. Zhang, Z. Gao, S. Li, L.-H. Chen, K. Liu, R. Cheng, X. Lin, J. Liu, Z. Li, J. Feng *et al.*, "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*, 2026, pp. 6664–6674.

[15] Z. Fan, M. Parelli, M. E. Kadoglou, X. Chen, M. Kocabas, M. J. Black, and O. Hilliges, "Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 494–504.

[16] W. Qu, Z. Cui, Y. Zhang, C. Meng, C. Ma, X. Deng, and H. Wang, "Novel-view synthesis and pose estimation for hand-object interaction from sparse views," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2023, pp. 15 054–15 065.

[17] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, "Reconstructing hands in 3d with transformers," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 9826–9836.

[18] J. Zhang, J. Deng, C. Ma, and R. A. Potamias, "Hawor: World-space hand motion reconstruction from egocentric videos," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 1805–1815.

[19] Z. Yu, S. Zafeiriou, and T. Birdal, "Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 27 716–27 726.

[20] H. Chen, T. Dong, T. Wu, L. Wang, Y. Jangir, Y. Niu, Y. Ye, H. Bharadhwaj, Z. Erickson, and J. Ichnowski, "Dexterous manipulation policies from rgb human videos via 3d hand-object trajectory reconstruction," *arXiv preprint arXiv:2602.09013*, 2026.

[21] J. Ren, Z. Jiang, Y. Yang, H. Jia, H. Jiang, B. Li, Y. Yao, C. Lin, Q. Shen, Z. Bing *et al.*, "C2dex: Contact-consistent reconstruction and retargeting for dexterous manipulation from monocular video," *arXiv preprint arXiv:2608.07045*, 2026.

[22] Q. Li, Y. Deng, Y. Liang, L. Luo, L. Zhou, C. Yao, L. Zeng, Z. Feng, H. Liang, S. Xu *et al.*, "Scalable vision-language-action model pre-training for robotic manipulation with real-life human activity videos," *arXiv preprint arXiv:2510.21571*, 2025.

[23] H. Bi, L. Wu, T. Lin, H. Tan, Z. Su, H. Su, and J. Zhu, "H-rdt: Human manipulation enhanced bimanual robotic manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 40, no. 22, 2026, pp. 18 135–18 143.

[24] R. Hoque, P. Huang, D. Yoon, J. Zhang *et al.*, "Egodex: Learning dexterous manipulation from large-scale egocentric video," in *International Conference on Learning Representations*, vol. 2026, 2026, pp. 4218–4237.

[25] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021, pp. 9040–9049.

[26] P. Banerjee, S. Shkodrani, P. Moulon, S. Hampali, S. Han, F. Zhang, L. Zhang, J. Fountain, E. Miller, S. Basol *et al.*, "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2025, pp. 7061–7071.

[27] X. Zhan, L. Yang, Y. Zhao, K. Mao, H. Xu, Z. Lin, K. Li, and C. Lu, "Oakink2: A dataset of bimanual hands-object manipulation in complex task completion," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2024, pp. 445–456.

[28] T. Kwon, B. Tekin, J. Stühmer, F. Bogo, and M. Pollefeys, "H2o: Two hands manipulating objects for first person interaction recognition," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 118–10 128, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233346960

[29] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.

[30] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 5026–5033.

[31] Z. Chen, S. Chen, E. Arlaud, I. Laptev, and C. Schmid, "Vividex: Learning vision-based dexterous manipulation from human videos," *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3336–3343, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:269330215

[32] Y. Xu, X. Mao, E. Miller, X. Yi, Y. Li, Z. Li, and R. B. Fisher, "Demobot: Efficient learning of bimanual manipulation with dexterous hands from third-person human videos," *ArXiv*, vol. abs/2601.01651, 2026.

[33] K. Ye, Y. Wu, S. Hu, J. Li, M. Liu, Y. Chen, and R. Huang, "Gen2real: Towards demo-free dexterous manipulation by harnessing generated video," *ArXiv*, vol. abs/2509.14178, 2025.

[34] W.-M. Zhou, B. Jiang, F. Yang, C. Paxton, and D. Held, "Learning hybrid actor-critic maps for 6d non-prehensile manipulation," *ArXiv*, vol. abs/2305.03942, 2023.

[35] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," in *Conference on Robot Learning*, 2022.

[36] P. Dan, K. Kedia, A. Chao, E. W. Duan, M. A. Pace, W.-C. Ma, and S. Choudhury, "X-sim: Cross-embodiment learning via real-to-sim-to-real," *ArXiv*, vol. abs/2505.07096, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:278501460

[37] T. G. W. Lum, O. Y. Lee, C. K. Liu, and J. Bohg, "Crossing the human-robot embodiment gap with sim-to-real rl using one human demonstration," *ArXiv*, vol. abs/2504.12609, 2025. [Online]. Available: https://api.semanticscholar.org/CorpusID:277857482

[38] X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, A. Lin, J.-W. Liu, Z. Ma, A. Sagar, B. Song, X. Wang, J. Yang, B. Zhang, P. Dollár, G. Gkioxari, M. Feiszli, and J. Malik, "SAM 3D: 3dfy anything in images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 7220–7232.

[39] D. S. Jung and K. M. Lee, "Learning dense hand contact estimation from imbalanced data," in *Advances in Neural Information Processing Systems*, vol. 38, 2025.

[40] B. Wen, W. Yang, J. Kautz, and S. T. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 868–17 879, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:266191252

[41] Y. Qin, W. Yang, B. Huang, K. V. Wyk, H. Su, X. Wang, Y.-W. Chao, and D. Fox, "Anyteleop: A general vision-based dexterous robot arm-hand teleoperation system," *ArXiv*, vol. abs/2307.04577, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259367735

[42] Y. Fu, Q. Feng, N. Chen, Z. Zhou, M. Liu, M. Wu, T. Chen, S. Rong, J. Liu, H. Dong *et al.*, "Cordvip: Correspondence-based visuomotor policy for dexterous manipulation in real-world," *arXiv preprint arXiv:2502.08449*, 2025.

[43] C. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 77–85, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:5115938

[44] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen, T. Liu, L. Yi, and H. Wang, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4737–4746,

2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:257280140

[45] J. Singla, A. Agarwal, and D. Pathak, "SAPG: Split and aggregate policy gradients," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 45 759–45 772.