

WristMimic: Full-Body Humanoid Control with Wrist-Guided Manipulation

Wongyun Yu¹, Youngwoon Kim¹, and Minsu Cho^{1,2}

¹ Pohang University of Science and Technology (POSTECH)

² RLWRLD

<https://wongyun-yu.github.io/wristmimic/>

Abstract. Retargeting human–object interaction demonstrations to physics-based simulation requires reproducing not only body motion but also the object motion and contacts that make manipulation succeed. However, position-only hand trajectories do not specify the contact forces needed to manipulate objects, and directly tracking them can overconstrain contact-rich finger behavior. We introduce WristMimic, a wrist-guided whole-body control framework that explicitly separates contact-free body motion from contact-rich hand manipulation. The contact-free body and wrist are guided by kinematic pose targets, whereas the fingers are not directly supervised by human hand pose. Instead, they learn grasping and manipulation behaviors from object tracking and contact outcomes. Our key insight is that the wrist is the natural gate between these two regimes. It is largely free from contact and can be tracked kinematically, yet it determines the global hand configuration and places the fingers within reachable grasp affordances. To ensure reliable wrist placement during interaction, we introduce wrist-specific reset constraints and reward prioritization. Experiments show that WristMimic matches or surpasses methods using full finger pose supervision while enabling finger-agnostic retargeting across diverse hand embodiments.

Keywords: Whole body control · Humanoid manipulation · Physics-based retargeting

1 Introduction

Retargeting human-object interaction demonstrations to physics-based simulation is a long-standing problem in robotics and computer vision [10, 12, 15, 20, 28, 72]. The goal of such retargeting extends beyond replicating body poses. Successful manipulation requires recovering correct object kinematics and contact dynamics in the target environment [9]. A common approach tracks kinematic trajectories from human demonstrations and replays them in simulation, which is effective for body parts that do not directly interact with the manipulated object, such as the torso, arms during reaching, and the wrist. However, for the hands, where most object interaction occurs, simply replaying kinematic trajectories is insufficient (Fig.1, left). For hands, the objective is not merely to

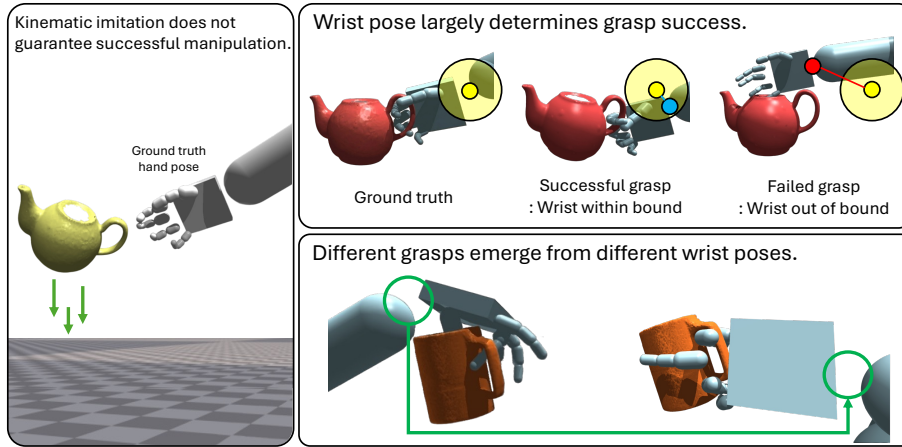


Fig. 1: Kinematic imitation does not guarantee successful manipulation and wrist pose encodes grasp affordance. **Left:** Even with accurate hand pose imitation, manipulation may fail because interaction forces are not encoded in kinematic trajectories. **Top right:** Wrist pose determines grasp success. Ground truth (left), feasible wrist region leading to stable grasp (middle, blue), and infeasible wrist pose causing failure (right, red). The yellow circle illustrates the feasible wrist boundary motivating our reset strategy. **Bottom right:** Different wrist poses induce distinct grasp configurations, highlighting that wrist pose captures grasp affordance and enables finger-agnostic, wrist-guided manipulation.

reproduce human finger poses, but to reproduce the object motion and contacts that realize manipulation. However, position-based kinematic data alone cannot encode the contact forces that govern such manipulation. This observation raises a natural question: rather than pursuing increasingly precise finger tracking, can we let the hands be guided by object and contact dynamics instead?

We address this question by decoupling whole-body control into two regimes: kinematic pose-guided control for contact-free body parts and object- and contact-guided control for contact-rich hand manipulation. This decoupling reflects a fundamental asymmetry in physics-based interaction. Body motion without direct object contact is often well-described in position space, while finger motion during interaction is dominated by contact force dynamics that kinematic trajectories cannot capture. Instead of relying on finger kinematic references, we use object kinematics as an outcome-level cue of successful control, together with contact objectives to shape physical interaction.

The key component that makes this decoupling possible is the wrist. The wrist is largely free from object contact, making it amenable to kinematic guidance. At the same time, it determines the global pose of the fingers and plays a key role in both affordance and grasp success. Consider grasping a cup by its handle versus its top (Fig.1, bottom right): the wrist position and orientation differ significantly between these two affordances, while finger configurations can

be learned from object-pose objectives and contact dynamics once the wrist is appropriately positioned.

Building on this wrist-centric insight, we introduce **WristMimic**, a whole-body control framework for affordance-aware manipulation that requires only body and wrist supervision. WristMimic positions the wrist in affordance-aware poses through wrist-specific reset boundaries and reward design, while allowing reinforcement learning to discover finger configurations through physics-based exploration, object-pose objectives, and contact dynamics without explicit finger supervision. As a consequence, this approach avoids relying on explicit finger pose supervision or complex finger-level capture data [4, 6, 14, 16, 25, 40–42, 53, 68, 73, 78]. Through extensive experiments, we demonstrate that precise wrist control achieves performance comparable to existing methods without relying on finger kinematic references, while supporting finger-agnostic retargeting across different hand embodiments. Our main contributions are:

- We propose a decoupled whole-body control framework that separates kinematic pose-guided body control from object- and contact-guided finger manipulation for human-object interaction retargeting.
- We identify the wrist as a critical bridge between control of contact-free body parts and contact-rich hand manipulation, and introduce wrist-specific reset boundaries and reward design to enable affordance-aware manipulation.
- We demonstrate that precise wrist control without finger kinematics achieves performance comparable to methods using full finger pose supervision, while supporting finger-agnostic retargeting across different hand embodiments.

2 Related Work

2.1 Physics-based Motion Imitation

Locomotion. Physics-based motion imitation has progressed from single-clip tracking [46] to large-scale whole-body control. AMP [48] and ASE [47] introduced adversarial motion priors for stylized control without explicit reference tracking. Scaling to large motion datasets, ScaDiver [62] employed mixture of experts, while PHC [34] achieved high-fidelity imitation on AMASS [37] dataset without external forces. PULSE [35] further distilled these motor skills into a universal latent representation reusable across downstream tasks. Recent universal trackers [11, 31, 69, 75] further improved generalization to unseen motions, and SONIC [36] demonstrated favorable scaling properties across data, model capacity, and compute. These works show that learning from kinematic trajectory is increasingly effective for contact-free body motion.

Manipulation. While locomotion imitation has been increasingly successful, extending physics-based imitation to manipulation remains challenging due to contact-rich dynamics and the difficulty of recovering interaction forces from kinematic data. Early approaches such as InterPhys [18] required task-specific reward design, limiting scalability. PhysHOI [5] introduced contact graphs for task-agnostic HOI imitation, and SkillMimic [58, 71] proposed skill-level imitation

robust to noisy demonstrations. Scaling further, InterMimic [66] achieved universal whole-body HOI control through teacher-student distillation with contact-guided rewards, while OmniGrasp [33] leveraged a universal dexterous motion representation to grasp over 1200 objects without paired full-body MoCap data. TokenHSI [44] and CLoSD [55] unified diverse interactions through task tokenization and diffusion-physics coupling respectively, and Wu *et al.* [63] proposed a complete pipeline from language instructions to physically plausible finger-object interactions. At the hand level, Chen *et al.* [9] propose a hierarchical wrist-guided framework for object-centric dexterous manipulation from human motion data. GraspXL [74] further studies large-scale grasp motion generation by learning policies that produce object-conditioned grasping trajectories across diverse objects and hand configurations. In contrast, we address whole-body HOI retargeting with decoupled supervision in a single policy. DexMachina [39], ManipTrans [29], and SPIDER [43] further advance dexterous retargeting through object-state tracking, residual transfer learning, and physics-informed embodiment bridging, respectively. Despite these advances, existing approaches either rely on dense kinematic supervision of detailed finger motion or focus on hand-level manipulation without integrating full-body control.

2.2 Dexterous Manipulation with Multi-Fingered Hands

Grasp Pose Generation. Early learning-based dexterous grasp generation benefited from differentiable contact modeling. Grasp’d [56] synthesizes multi-finger grasps by optimizing through differentiable contact-rich simulation. The availability of large-scale synthetic grasp datasets such as DexGraspNet [57] shifted the field toward data-driven prediction of stable grasp poses. Subsequent works improved coverage and multimodality with proposal-and-policy formulations [67] and generative modeling, including diffusion and transformer-based grasp synthesis [32, 61, 64]. Beyond purely geometric conditioning, recent methods incorporate task intent via language. DexGYS [59] enables language-guided dexterous grasp generation, while SemGrasp [30] and DexVLG [19] further align grasp representations with vision-language priors. Dexonomy [7] organizes grasps into a taxonomy and AffordDexGrasp [60] targets open-set language-guided affordances. DextER [26] introduces contact-based embodied reasoning by predicting which finger links should contact where on the object surface before generating the final grasp, improving intention alignment and enabling steerable grasp synthesis. While these methods generate text and affordance-aligned grasp poses, they focus primarily on the final grasp configuration, rather than learning physics-aware trajectories or accurate manipulation sequences.

Dexterous Manipulation Learning. Beyond static grasp pose generation, dexterous manipulation learning targets temporally extended, closed-loop control for contact-rich multi-fingered hands. Early successes with large-scale reinforcement learning demonstrated agile in-hand manipulation via massive simulation and domain randomization [1, 2]. Imitation learning offers an alternative supervision source: Qin *et al.* [50] transfers single-camera teleoperation to multiple hand embodiments, while EgoMimic [23] scales policy learning from egocen-

tric human videos. Subsequent work improved robustness and horizon length, including DeXtreme [17] for sim-to-real agile in-hand skills and Sequential Dexterity [8] for chaining dexterous policies over long-horizon tasks. To better leverage contact information, RotateIt [49] integrates vision and tactile for general in-hand rotation. Recently, foundation-model-driven policies have begun to couple task-level reasoning with low-level dexterity: DexGraspVLA [77] combines a VLM-based planner with a diffusion action controller trained by imitation for robust, instruction-following dexterous grasping, and DexUMI [65] explores using human hands as a universal manipulation interface for scalable demonstration collection. EgoScale [76] shows that scaling egocentric human data can serve as predictable supervision for dexterous manipulation by pretraining on large-scale wrist and retargeted hand actions and transferring to real robots with minimal robot data. These hand-centric advances motivate our study of how to integrate dexterous manipulation into whole-body retargeting while reducing reliance on finger kinematic references.

3 Method

Overview. We train a policy network that outputs actions in a physics-based simulator to reproduce a reference human-object interaction sequence. The reproduced motion must (i) track the reference body trajectory and (ii) manipulate the object so that its motion follows the reference kinematic trajectory. For the first one, standard imitation learning is often sufficient as the objective primarily involves joint movement. However, accurate object manipulation cannot be achieved by body pose imitation alone. Motion capture is imperfect, and position-based trajectories do not encode interaction forces; for example, grasping an object with weak or strong force may correspond to identical finger positions. As a result, correct pose imitation does not necessarily imply correct manipulation. Since most manipulation-related contacts occur at the hands, we decouple supervision by contact regime: the policy tracks kinematic pose targets only for object-contact-free joints, including the body and wrists, while finger joints receive no kinematic targets and are instead learned through object motion and contact outcomes, which reflect successful physical interaction. To make this decoupled, object-driven learning formulation reliable, we treat the wrist as the bridge between contact-free body control and contact-rich manipulation. We constrain wrist behavior around contact using a contact-window with time-varying reward weights and phase-specific reset thresholds. Our pipeline is illustrated in Fig.2 and detailed in Alg.1.

3.1 Problem Formulation.

Given reference motion data $\{\hat{\mathbf{q}}_t, \hat{\mathbf{q}}_t^{\text{obj}}, \hat{\mathbf{c}}_t\}_{t=1}^T$ including body poses $\hat{\mathbf{q}}_t$, object trajectories $\hat{\mathbf{q}}_t^{\text{obj}}$, and contact labels $\hat{\mathbf{c}}_t$, we learn a physics-based policy π which outputs action $\mathbf{a}_t$ to generate simulated states $\{\mathbf{q}_t, \mathbf{q}_t^{\text{obj}}, \mathbf{c}_t\}$. We formulate this as a MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$ with state space $\mathcal{S}$, action space $\mathcal{A}$, simulator transition dynamics $\mathcal{T}$, reward $\mathcal{R}$, and discount factor γ .

The state $\mathbf{s}_t = (\mathbf{s}_t^p, \mathbf{s}_t^g)$ consists of a proprioceptive state $\mathbf{s}_t^p$ from simulator and a goal state $\mathbf{s}_t^g$ from the reference. The policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$ predicts actions for all 51 joints, with actions sampled from Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_t, \sigma^2 \mathbf{I})$. Here $\boldsymbol{\mu}_t \in \mathbb{R}^{51 \times 3}$ are predicted means and σ is fixed standard deviation. We optimize π with PPO [51] to maximize reward $r_t = \mathcal{R}(\mathbf{s}_t^p, \mathbf{s}_t^g)$ measuring alignment between the simulated state and the reference. Detailed policy representation is in §3.2.

3.2 Policy Representation for Decoupled Manipulation

Proprioceptive State. The humanoid follows the SMPL-X [45] configuration with 52 rigid bodies connected by 51 actuated joints, each with 3 DoFs. The 51 actuated joints are partitioned into two sets: $\mathcal{J}_{cf}$ (object-contact-free joints) consisting of 19 body joints and 2 wrists, and $\mathcal{J}_{cr}$ (contact-rich joints) consisting of 30 finger joints. We denote the root pose as $\mathbf{q}_t^{\text{root}} = (\mathbf{p}_t^{\text{root}}, \boldsymbol{\theta}_t^{\text{root}}) \in \mathbb{R}^3 \times \mathbb{R}^6$ and velocity as $\dot{\mathbf{q}}_t^{\text{root}} = (\mathbf{v}_t^{\text{root}}, \boldsymbol{\omega}_t^{\text{root}}) \in \mathbb{R}^3 \times \mathbb{R}^3$, where $\mathbf{p}$, $\boldsymbol{\theta}$, $\mathbf{v}$, and $\boldsymbol{\omega}$ represent position, orientation, linear velocity, and angular velocity, respectively. The 51 actuated joints have poses $\mathbf{q}_t \in \mathbb{R}^{51 \times 3} \times \mathbb{R}^{51 \times 6}$ and velocities $\dot{\mathbf{q}}_t \in \mathbb{R}^{51 \times 3} \times \mathbb{R}^{51 \times 3}$ with the same representation. The 52 rigid bodies are used to compute body-level contacts and interaction graph features. Object state is represented by $\mathbf{q}_t^{\text{obj}}$ with velocities $\dot{\mathbf{q}}_t^{\text{obj}}$. Binary contact markers $\mathbf{c}_t \in \mathbb{R}^{52}$ indicate contact for each rigid body, and interaction graph vectors $\mathbf{i}_t \in \mathbb{R}^{52 \times 3}$ encode the spatial relationship between each humanoid rigid body and its nearest object surface point [13, 22, 52, 70]. The proprioceptive state is:

$$\mathbf{s}_t^p = \{\mathbf{q}_t^{\text{root}}, \dot{\mathbf{q}}_t^{\text{root}}, \mathbf{q}_t, \dot{\mathbf{q}}_t, \mathbf{q}_t^{\text{obj}}, \dot{\mathbf{q}}_t^{\text{obj}}, \mathbf{c}_t, \mathbf{i}_t\}. \quad (1)$$

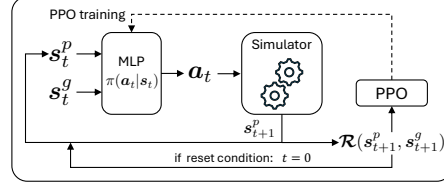


Fig. 2: Training pipeline. The policy observes proprioceptive and goal state to generate actions. Rewards measure alignment with reference motion using time-varying weights that prioritize contacts and object dynamics. Alg.1 details the complete training procedure.

Algorithm 1: WristMimic Training

```

1: Input: Reference  $\{\hat{\mathbf{q}}_t, \hat{\mathbf{q}}_t^{\text{obj}}, \hat{\mathbf{c}}_t\}$ , contact frame  $t_c$ 
2: Preprocessing: Identify grasping hands  $\mathcal{J}_{\text{arm}}$ 
3: for episode = 1 to  $N_{\text{episodes}}$  do
4:   Sample  $\mathbf{s}_0$  from reference
5:   while  $t < T_{\text{max}}$  and not terminated do
6:      $\mathbf{s}_t^p \leftarrow \text{get\_proprioception}()$ 
7:      $\mathbf{s}_t^g \leftarrow \text{get\_goal}()$  (excludes 30 finger joints)
8:      $\mathbf{a}_t \sim \pi(\cdot | \mathbf{s}_t^p, \mathbf{s}_t^g)$  (actions for 51 joints)
9:      $\mathbf{s}_{t+1}^p \leftarrow \text{sim.step}(\mathbf{a}_t)$ 
10:    Compute time-varying weights:
11:    
$$w_j(t) = \begin{cases} 0 & j \in \mathcal{J}_{\text{arm}}, t \in [t_c - t_b, t_c + t_a] \\ 1 & j \in \mathcal{J}_{\text{wrist}}, t \in [t_c - t_b, t_c + t_a] \\ w_{\text{red}} & j \in \mathcal{J}_{\text{cf}} \setminus (\mathcal{J}_{\text{arm}} \cup \mathcal{J}_{\text{wrist}}), \\ & t \in [t_c - t_b, t_c + t_a] \\ 1 & \text{otherwise} \end{cases}$$

12:    Compute reward components:
13:     $r_t^a = \exp(-\lambda^a \sum_j w_j(t) E_j^a)$  where  $a \in \{p, \theta\}$ 
14:     $r_t^{a,\text{obj}} = \exp(-\lambda^{a,\text{obj}} E^{a,\text{obj}})$  where  $a \in \{p, \theta\}$ 
15:     $r_t^c = \exp(-\lambda^c E^c)$ 
16:    Compute multiplicative reward:
17:     $r_t = r_t^p \cdot r_t^\theta \cdot r_t^{p,\text{obj}} \cdot r_t^{\theta,\text{obj}} \cdot r_t^c$ 
18:    if base termination conditions met then
19:      terminated  $\leftarrow$  True
20:    end if
21:    if  $t \in [t_c - t_b, t_c + t_a]$  then
22:      phase  $\leftarrow \text{get\_phase}(t, t_c)$ 
23:      if wrist error > phase-specific threshold then
24:        terminated  $\leftarrow$  True
25:      end if
26:    end if
27:    Store  $(\mathbf{s}_t, \mathbf{a}_t, r_t)$ 
28:  end while
29:  Update  $\pi$  and  $V$  via PPO when buffer full
30: end for
```

Goal State. The goal state $\mathbf{s}_t^g = \{g_{t,t+k}\}_{k \in K}$ encodes differences between the current state and future reference states at time horizons K :

$$g_{t,t+k} = \{\Delta\boldsymbol{\theta}_{t+k}, \Delta\mathbf{p}_{t+k}, \Delta\boldsymbol{\theta}_{t+k}^{\text{obj}}, \Delta\mathbf{p}_{t+k}^{\text{obj}}, \Delta\mathbf{i}_{t+k}, \Delta\mathbf{c}_{t+k}\}, \quad (2)$$

where Δ denotes the difference between the reference state at k frames ahead and the current state, using $\ominus$ for rotation differences and $-$ for others. For decoupled supervision, we assign kinematic pose targets only to joints in $\mathcal{J}_{\text{cf}}$ and exclude all joints in $\mathcal{J}_{\text{cr}}$. No kinematic reference trajectories are provided for finger joints. As a result, finger behavior is not supervised through pose imitation, but is instead optimized indirectly through object motion and contact objectives, *i.e.*, via the object pose targets $\Delta\boldsymbol{\theta}_{t+k}^{\text{obj}}$ and $\Delta\mathbf{p}_{t+k}^{\text{obj}}$, together with contact alignment rewards. The contact target $\Delta\mathbf{c}_{t+k} \in \mathbb{R}^{23}$ consists of 21 explicit contact labels for the non-finger rigid bodies and 2 binary hand-level contact indicators (one per hand). A hand-level contact indicator is set to 1 if any rigid body belonging to that hand is in contact with the object. All goal elements are normalized relative to the root position and facing direction [3, 33–35, 54, 58, 66, 71].

Reward. Our reward function follows a multiplicative structure:

$$r_t = r_t^p \cdot r_t^\theta \cdot r_t^{p,\text{obj}} \cdot r_t^{\theta,\text{obj}} \cdot r_t^c, \quad (3)$$

where each component takes the form $r = \exp(-\lambda E)$, penalizing error E between simulated and reference states with component-specific weights λ . The five components measure: joint position tracking (r_t^p), joint rotation tracking (r_t^θ), object position tracking ($r_t^{p,\text{obj}}$), object rotation tracking ($r_t^{\theta,\text{obj}}$), and contact alignment (r_t^c). Following InterMimic [66], E^p and E^θ prioritize rotation when joints are far from the object and position when joints are near.

To implement decoupled supervision, we compute pose-tracking errors (r_t^p , r_t^θ) only over joints in $\mathcal{J}_{\text{cf}}$, and exclude all 30 finger joints from direct kinematic tracking. For contact alignment, we compute error over the 23 contact elements defined in the goal state (21 non-finger bodies + 2 hand-level indicators). As a result, contact-free body parts receive reward signals from kinematic references, while finger joints are influenced by object pose tracking and contact alignment. This reward structure aligns with our decoupled formulation, where finger joints are not directly supervised by pose targets but are shaped by object motion and contact alignment.

Action. Actions $\mathbf{a}_t \in \mathbb{R}^{51 \times 3}$ specify target rotations for all 51 actuated joints in exponential map coordinates [66], which are converted to torques via PD controllers. Although finger joints receive no kinematic supervision in the goal state, they remain fully actuated in the action space. Their behavior is learned implicitly through object pose rewards, contact alignment, and interaction dynamics.

3.3 Learning Wrist-Guided Manipulation

After decoupling kinematic supervision for contact-free body parts from object- and contact-guided finger behavior, we identify the wrist as the structural bridge

between these two regimes. While finger behavior is learned through object motion and contact outcomes, effective exploration requires the hand to be placed in an affordance-consistent and reachable configuration. Since the wrist determines the global pose of the fingers and the reachable contact set, a mis-aligned wrist may prevent access to the intended contact region. Moreover, as the wrist is typically free from direct object contact, it remains suitable for kinematic guidance using reference trajectories. We therefore prioritize wrist placement through selective relaxation of proximal arm supervision and wrist-specific phase-dependent reset thresholds. Finger configurations are then allowed to emerge from interaction dynamics.

Prioritized Alignment Strategy. To ensure feasible wrist placement that enables finger behavior to emerge through interaction dynamics, we adopt a prioritized alignment strategy. Motion capture data is inherently imperfect when executed in physics-based simulation, making it difficult to simultaneously align all body joints with the reference motion. For example, motion capture sequences are often defined with a floating root frame that may violate physical constraints in simulation. As a result, some joints must be allowed to deviate to achieve accurate wrist alignment. We define a *contact window* $[t_c - t_b, t_c + t_a]$ centered on the first contact frame t_c , where t_b and t_a denote frames before and after contact. Within this window, we enforce tight wrist constraints while relaxing upper arm supervision, allowing the policy to adjust upper arm configurations freely to position the wrist optimally. This strategy is implemented through two complementary mechanisms: reward weight modulation and phase-specific reset boundaries.

Reward Weight Modulation. To prioritize wrist alignment while relaxing upper arm supervision, we modulate joint reward weights over time. Specifically, the joint rewards are formulated to exclude all 30 finger joints:

$$r_t^\alpha = \exp \left(-\lambda^\alpha \sum_{j \in \mathcal{J}_{cf}} w_j(t) E_j^\alpha \right) \quad \text{where } \alpha \in \{p, \theta\}, \quad (4)$$

where $\mathcal{J}_{cf}$ denotes the 19 body joints and 2 wrist joints, and $w_j(t)$ are time-varying weight factors defined as:

$$w_j(t) = \begin{cases} 0, & \text{if } j \in \mathcal{J}_{arm}, t \in [t_c - t_b, t_c + t_a], \\ 1, & \text{if } j \in \mathcal{J}_{wrist}, t \in [t_c - t_b, t_c + t_a], \\ w_{red}, & \text{if } j \in \mathcal{J}_{cf} \setminus (\mathcal{J}_{arm} \cup \mathcal{J}_{wrist}), t \in [t_c - t_b, t_c + t_a], \\ 1, & \text{otherwise.} \end{cases} \quad (5)$$

where $\mathcal{J}_{arm}$ denotes the shoulder and elbow joints of the grasping hand(s), $\mathcal{J}_{wrist}$ denotes the wrist joints, and $w_{red} < 1$ is the reduced weight factor. This selective relaxation removes proximal arm constraints ($w_j = 0$), keeps wrist supervision unchanged ($w_j = 1$), and reduces the weights of the remaining contact-free body joints ($w_j = w_{red}$) during the contact window. Moreover, reducing other body joint weights ($w_j = w_{red}$) allows the policy to focus on hand-object contact and

object dynamics in contact window. Specific weight values are detailed in the supplementary material.

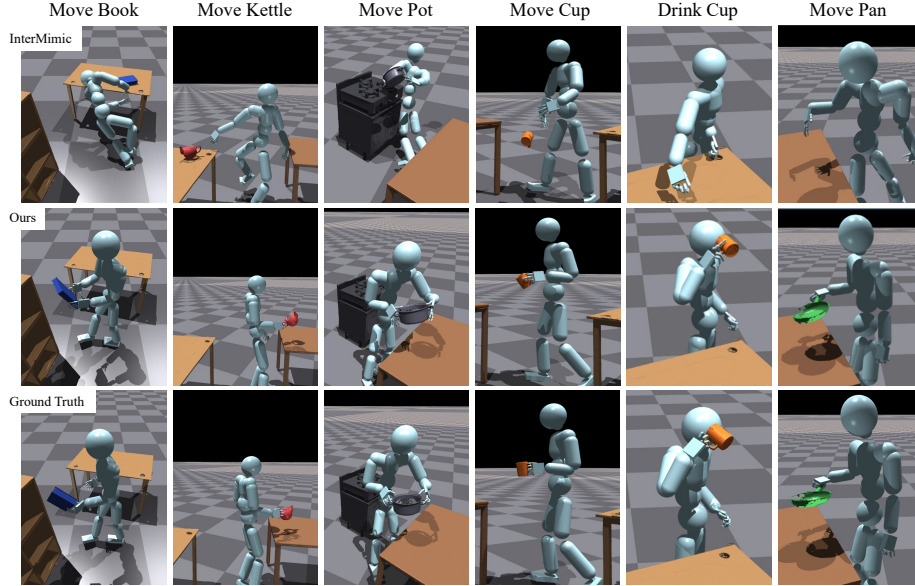


Fig. 3: Qualitative comparison across 6 representative scenes. We illustrate representative qualitative results comparing our method with InterMimic [66]. Our method achieves affordance-aware grasping with diverse manipulation behaviors while accurately reproducing whole-body motion in the physics-based simulation.

Phase-Specific Reset Thresholds. We employ early termination to prevent inefficient exploration when episodes diverge from feasible trajectories. Specifically, episodes terminate if: **(i)** body pose deviates severely from reference, **(ii)** root height drops below a threshold (indicating fall), **(iii)** object pose drifts too far from reference, or **(iv)** contact error persists for multiple frames [33, 66]. In addition to this base reset strategy, we apply phase-specific reset thresholds to the wrist of grasping hand(s) within the contact window. We subdivide the contact window into three phases centered on t_c : *approach* $[t_c - t_b, t_c + \tau_1]$, *grasping* $[t_c + \tau_1, t_c + \tau_2]$, and *stabilization* $[t_c + \tau_2, t_c + t_a]$, where $0 \leq \tau_1 < \tau_2 \leq t_a$ define phase transitions after first contact. During the *grasping* phase, we apply tight thresholds of 7cm in position and 0.2 radians in orientation, terminating if wrist deviates beyond these bounds from reference (Fig.1, top right). The *approach* and *stabilization* phases use looser thresholds of 15cm and 0.5 radians, allowing trajectory exploration while maintaining approximate alignment. This design—using reduced reward weights with strict reset boundaries rather than high reward weights—enables the policy to discover optimal wrist poses within a feasible search space, accommodating mocap inaccuracies while ensuring natural grasp configurations.

Method	Success rate (%) $\uparrow$ / Obj. pos. err. (cm) $\downarrow$ / Obj. rot. err. ($^\circ$) $\downarrow$					
	OMOMO [27]		ParaHome [24]		Avg.	
InterMimic [66]	86.6	/ 14.2 / 22.2	0.1	/ 83.5 / 73.3	43.3	/ 48.8 / 47.7
SkillMimicV2 [71]	-		1.3	/ 72.1 / 80.3	1.3	/ 72.1 / 80.3
WristMimic (Ours)	98.9	/ 7.3 / 12.2	83.3	/ 15.3 / 33.9	91.1	/ 11.3 / 23.1

Table 1: Comparison with state-of-the-art methods. We compare our method against recent WBC-based object interaction baselines. Results are reported across 40 diverse interaction sequences sampled from OMOMO (20) [27], and ParaHome (20) [24]. Note that SkillMimicV2 [71] is evaluated only on ParaHome, as it specifically utilizes bone-vector representations. **Bold** indicates the best performance.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate WristMimic on two human-object interaction datasets: ParaHome [24] and OMOMO [27]. These datasets provide object and subject geometries, human motion trajectories represented via SMPL-X [45] or bone vectors, and corresponding object trajectories. Each episode spans approximately 100–250 frames. We train per-scene policies on 20 representative sequences from ParaHome and OMOMO respectively. This setup allows us to evaluate the robustness of our wrist-guided, object-kinematics-driven approach across diverse manipulation scenarios. We assess performance using three primary metrics across 10,000 rollout trials per scene:

- Success rate (%): A trial is successful if the sequence reaches completion without early termination, the average object position error remains below 10 cm, and the robot maintains contact for at least 80% of the reference contact duration.
- Object position/rotation error (cm, $^\circ$): Aligning with our principle that manipulation imitation should prioritize object trajectory over exact finger re-play, we measure the average position/rotation deviation of the object from its reference trajectory throughout the sequence.

Implementation Details. We implement WristMimic in Isaac Gym [38] and train all policies using Proximal Policy Optimization (PPO) [51]. Training uses 2048 parallel environments on a single NVIDIA RTX 3090 GPU with 30 Hz control frequency. The policy network consists of separate actor and critic MLPs with SiLU [21] activations. Contact window parameters are set to $t_b = 10$, $t_a = 15$ frames, with phase transitions at $\tau_1 = -2$ and $\tau_2 = 12$ frames relative to the first contact frame t_c . Complete hyperparameters are provided in the supplementary material.

4.2 Quantitative Results

We compare WristMimic against InterMimic [66] and SkillMimicV2 [71]. For fair comparison, we train per-scene policies for all methods. Evaluation is conducted on 40 sequences in total, consisting of 20 sequences from ParaHome [24]

and 20 from OMOMO [27]. For SkillMimicV2, evaluation is performed only on ParaHome, since SkillMimicV2 requires bone-vector representations, which are available only in ParaHome.

Quantitative results are shown in Tab.1. WristMimic achieves the best performance across all evaluation metrics despite not relying on explicit finger pose supervision. In particular, on the OMOMO dataset [27], WristMimic reduces both object position and rotation errors by nearly half compared to recent methods. These results indicate that guiding manipulation through wrist placement and object kinematics provides a strong structural signal for object-trajectory-aware control.

On the ParaHome dataset [24], both InterMimic and SkillMimicV2 exhibit significantly lower success rates. Notably, while InterMimic shows a clear performance gap between OMOMO and ParaHome, WristMimic maintains relatively consistent performance across the two datasets. Since ParaHome contains more dexterous interaction scenarios such as grasping small handles or objects with narrow contact regions, these results suggest that wrist-guided constraints provide robust structural guidance for delicate grasping.

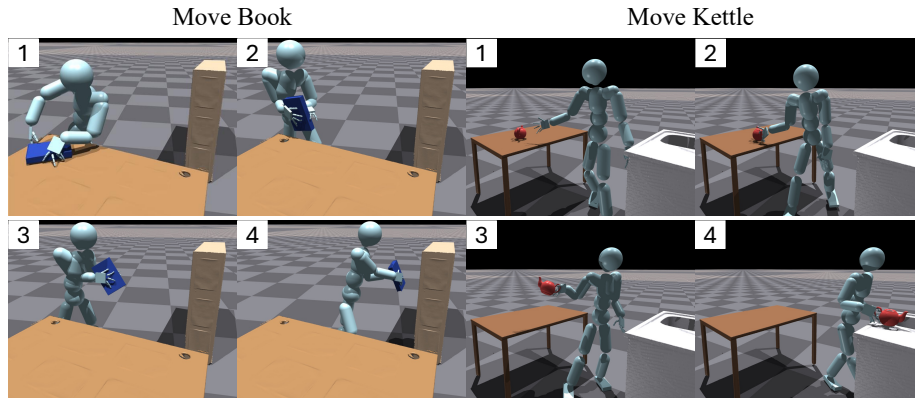


Fig. 4: Key frames from manipulation sequences. Our method successfully handles challenging manipulation scenarios, including grasping a book from a flat table without a handle and grasping a kettle followed by turning the body while maintaining a stable grasp.

4.3 Qualitative Results

We present representative qualitative results comparing our method with InterMimic [66] on six scenes from the ParaHome dataset (Fig.3). As discussed in §4.2, previous methods often struggle in these dexterous manipulation scenarios. In particular, as illustrated in Fig.5, even when SkillMimicV2 produces visually plausible finger configurations, incorrect wrist placement prevents stable object grasping. This example highlights that correct wrist placement is a prerequisite

Method	Success rate(%) $\uparrow$ / Obj. pos. err.(cm) $\downarrow$ / Obj. rot. err.($^\circ$) $\downarrow$
	ParaHome [24]
Decoupled (No Wrist Constraints)	0.0 / 56.4 / 49.7
+ reward weight modulation only	0.0 / 50.3 / 43.6
+ phase specific reset only	67.5 / 28.8 / 56.9
WristMimic (Ours)	86.5 / 9.9 / 36.0

Table 2: Ablation on wrist constraints. We evaluate the effect of removing wrist-specific constraints on the ParaHome dataset. WristMimic significantly improves both success rate and object pose accuracy compared to the decoupled formulation without wrist constraints, highlighting the importance of wrist guidance for stable object manipulation. **Bold** indicates the best performance.

for establishing feasible contact configurations for grasping. In contrast, our approach first establishes a feasible wrist configuration, allowing finger poses to naturally adapt to the object and enabling both stable grasping and accurate tracking of the reference object trajectory. These examples demonstrate that guiding manipulation through wrist placement enables affordance-consistent grasping in dexterous interaction scenarios.

Fig.4 further shows sequential frames sampled at 50-frame intervals from two representative manipulation sequences. WristMimic maintains stable object contact while producing natural hand poses throughout the sequence, demonstrating consistent manipulation behavior over time.

In the Move Book scenes in Fig.3 and Fig.4, a slight deviation from the ground truth motion can be observed. The human demonstration initially lifts the book using both hands before releasing one hand, whereas our policy continues to support the book with both hands due to grasping stability. This behavior reflects our method’s design choice to prioritize stable object manipulation and trajectory tracking over strict imitation of the human body pose.

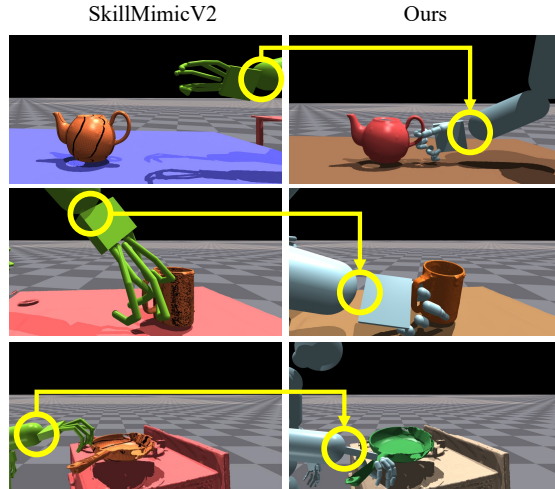


Fig. 5: Grasping pose comparison. Although SkillMimicV2 demonstrates plausible finger closure, the incorrect wrist pose prevents a successful grasp, whereas WristMimic maintains precise wrist alignment.

4.4 Ablation Study

Ablation on Wrist Constraints. We evaluate the effect of wrist-specific constraints on 8 representative ParaHome [24] sequences. As shown in Tab.2, the

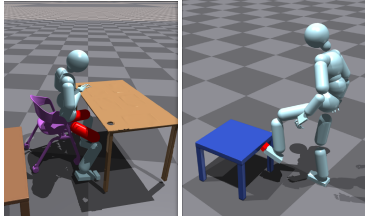


Figure 6. Scenes involving non-hand body contact.

Scene	SR (%) $\uparrow$	Pos. err. (cm) $\downarrow$	Rot. err. ($^\circ$) $\downarrow$
Largetable	99.9	9.9	10.6
Whitechair	99.7	8.0	6.8
Sit on chair #1	97.2	9.8	5.0
Sit on chair #2	99.8	8.5	6.1

Table 3. Rows 1–2: pushing object with foot. Rows 3–4: sit on chair.

decoupled formulation without wrist constraints performs significantly worse, indicating that decoupling alone is not sufficient for reliable manipulation. We further decompose our wrist-centric design into two components: reward weight modulation within the contact window and phase-specific wrist reset thresholds. Phase-specific resets provide the main improvement by keeping exploration near feasible wrist poses, while reward weight modulation alone gives only marginal gains. However, combining reward modulation with phase-specific resets yields the best overall performance, substantially improving success rate and object pose accuracy. These results show that WristMimic’s gains come from explicitly guiding wrist placement during contact-rich phases, which is critical for stabilizing hand–object interactions and achieving affordance-consistent grasps.

Ablation on Reset Threshold. We further ablate the grasping-phase wrist reset thresholds on four representative ParaHome [24] scenes. As shown in Tab.4, 3.5 cm/0.1 rad (tight) is too restrictive for exploration and pre-grasp recovery, while 15 cm/0.5 rad (loose) admits implausible wrist states that destabilize contact. Our default 7 cm/0.2 rad achieves the best trade-off between feasible exploration and reliable wrist placement.

Ablation on sequences of non-hand-contact bodies.

Fig.6 and Tab.3 present ablation studies on non-hand body contact scenarios, where body parts other than the hands serve as the primary contact regions. Unlike

Variant	SR (%) $\uparrow$	Pos. err. (cm) $\downarrow$	Rot. err. ($^\circ$) $\downarrow$
(a) Tight	0.0	99.8	86.8
(b) Ours	96.3	9.5	18.1
(c) Loose	17.5	78.3	72.3

Table 4. Reset threshold ablation.

hand-object manipulation, these interactions mainly involve coarse body-level support or displacement, and thus can be handled by the standard whole-body mimicking objective without additional task-specific design. The results show that such non-hand contacts do not require any wrist-like design, allowing our framework to extend naturally to scenarios such as sitting on a chair and displacing a small table with the foot.

Hand Morphology Generalization. Tab.5 demonstrates that our finger-agnostic formulation enables learning across different hand morphologies without requiring finger-specific motion data. We retrain policies using three hand configurations: the default scene-specific SMPL-X hand, the hand model from InterMimic [66], and the OmniGrasp [33] hand, evaluated on 10 sequences from ParaHome [24]. These configurations differ substantially in hand size, joint lengths, and initial wrist and finger poses. Despite these differences, grasping remains

	Success rate(%) $\uparrow$ / Obj. pos. err.(cm) $\downarrow$ / Obj. rot. err.($^\circ$) $\downarrow$
Method	ParaHome [24]
InterMimic [66]	95.2 / 10.8 / 28.1
OmniGrasp [33]	76.4 / 16.9 / 27.1
Scene-specific	75.8 / 20.1 / 39.6

Table 5: Hand morphology generalization performance. Scene-specific denotes the default hand following the scene-specific SMPL-X [45] configuration. We evaluate the same policy across different hand morphologies, including the InterMimic and OmniGrasp hand models, which vary in hand size, joint length, and joint limits.

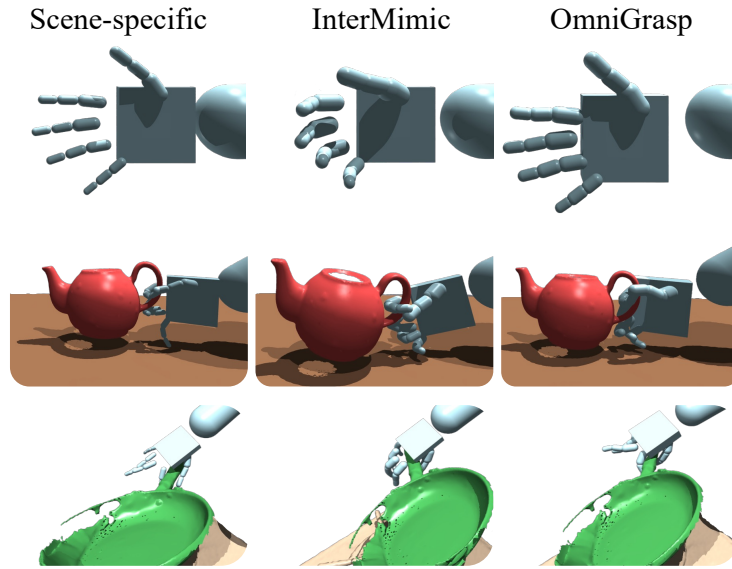


Fig. 7: Ablation on hand-generalization. Our finger-agnostic approach generalizes across different hand morphologies without finger-specific data. We demonstrate successful grasping with three different hands that differ in joint length, size, and pose.

successful across all configurations, suggesting that wrist guidance provides a strong structural signal for affordance-aware grasping.

Fig.7 visualizes representative successful manipulations with the three hand types. All policies are trained independently using the same wrist-guided strategy without explicit finger pose supervision. As shown in Tab.5, performance remains comparable across different hand morphologies, indicating that the policy adapts to varying hand geometries through interaction dynamics rather than kinematic supervision. Among the three configurations, the InterMimic [66] hand exhibits the most realistic finger poses and highest performance due to its constrained joint ranges that prevent physically implausible poses while other hands set joint range as $[-\pi, \pi]$. Since the joint range-constrained setting is more realistic in the real world, this shows that the wrist-only approach can make natural grasping poses emerge while maintaining grasping quality.

4.5 Limitations

Our method focuses on affordance-aware grasping and object manipulation rather than precise in-hand manipulation that requires detailed finger pose control. While wrist-guided learning enables the policy to discover natural grasp configurations through interaction dynamics, tasks involving fine-grained finger repositioning or object reorientation within the hand remain beyond the current scope. Extending the framework to support such in-hand manipulation without explicit finger supervision is an interesting direction for future work.

Another limitation is that our current framework trains scene-specific policies, requiring separate training for each scene. This contrasts with methods such as OmniGrasp [33], InterMimic [66], and MaskedManipulator [54], which aim to generalize across scenes. Nevertheless, our experiments show that the wrist-guided formulation generalizes across different hand morphologies within the same scene, suggesting that the learned interaction structure is not tied to a specific hand model. Developing scene-general policies that retain this wrist-centric interaction structure remains an important direction for future research.

5 Conclusion

We have presented WristMimic, a wrist-guided framework for physics-based re-targeting of human-object interaction demonstrations. Our approach decouples supervision between contact-free body motion and contact-rich hand manipulation, and highlights the wrist as a key bridge between these two regimes. Because the wrist largely remains contact-free while determining the global configuration of the fingers, guiding the wrist provides a useful structural signal for learning manipulation behaviors without explicit finger pose supervision. Through experiments across multiple datasets and hand morphologies, we show that wrist-guided learning can achieve grasping and object manipulation performance comparable to methods that rely on detailed finger supervision. These results suggest that emphasizing wrist guidance may offer a practical direction for reducing reliance on dense finger motion capture in physics-based manipulation learning. More broadly, this perspective suggests that manipulation policies can be learned by focusing on structurally important joints rather than densely supervising all degrees of freedom.

Acknowledgements

This work was supported by Samsung Electronics (IO201208-07822-01) and also by IITP grants (RS-2022-II220113: Developing a Sustainable Collaborative Multi-modal Lifelong Learning Framework (50%), RS-2022-II220290: Visual Intelligence for Space-Time Understanding & Generation (25%), RS-2024-00457882: National AI Research Lab Project (20%), RS-2019-III191906: AI Graduate School Program at POSTECH (5%)) funded by Ministry of Science and ICT, Korea.

References

1. Akkaya, I., Andrychowicz, M., Chociej, M., Litwin, M., McGrew, B., Petron, A., Paino, A., Plappert, M., Powell, G., Ribas, R., et al.: Solving rubik’s cube with a robot hand. arXiv preprint arXiv:1910.07113 (2019)
2. Andrychowicz, O.M., Baker, B., Chociej, M., Jozefowicz, R., McGrew, B., Pachocki, J., Petron, A., Plappert, M., Powell, G., Ray, A., et al.: Learning dexterous in-hand manipulation. *The International Journal of Robotics Research* **39**(1), 3–20 (2020)
3. Bae, J., Won, J., Lim, D., Min, C.H., Kim, Y.: Pmp: Learning to physically interact with environments using part-wise motion priors. In: *International Conference on Computer Graphics and Interactive Techniques* (2023)
4. Banerjee, P., Shkodrani, S., Moulon, P., Hampali, S., Han, S., Zhang, F., Zhang, L., Fountain, J., Miller, E., Basol, S., et al.: Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 7061–7071 (2025)
5. Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: *2024 International Conference on 3D Vision (3DV)*. pp. 464–473. IEEE (2024)
6. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., 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)
7. Chen, J., Ke, Y., Peng, L., Wang, H.: Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy. arXiv preprint arXiv:2504.18829 (2025)
8. Chen, Y., Wang, C., Fei-Fei, L., Liu, C.K.: Sequential dexterity: Chaining dexterous policies for long-horizon manipulation. arXiv preprint arXiv:2309.00987 (2023)
9. Chen, Y., Wang, C., Yang, Y., Liu, C.K.: Object-centric dexterous manipulation from human motion data. arXiv preprint arXiv:2411.04005 (2024)
10. Chen, Z., Chen, S., Arlaud, E., Laptev, I., Schmid, C.: Vividex: Learning vision-based dexterous manipulation from human videos. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 3336–3343. IEEE (2025)
11. Chen, Z., Ji, M., Cheng, X., Peng, X., Peng, X.B., Wang, X.: Gmt: General motion tracking for humanoid whole-body control. arXiv preprint arXiv:2506.14770 (2025)
12. Cheng, X., Ji, Y., Chen, J., Yang, R., Yang, G., Wang, X.: Expressive whole-body control for humanoid robots. arXiv preprint arXiv:2402.16796 (2024)
13. Dahiya, R.S., Metta, G., Valle, M., Sandini, G.: Tactile sensing—from humans to humanoids. *IEEE transactions on robotics* **26**(1), 1–20 (2009)
14. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: Arctic: A dataset for dexterous bimanual hand-object manipulation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12943–12954 (2023)
15. Fu, Z., Zhao, Q., Wu, Q., Wetzstein, G., Finn, C.: Humanplus: Humanoid shadowing and imitation from humans. arXiv preprint arXiv:2406.10454 (2024)
16. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: 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)
17. Handa, A., Allshire, A., Makoviychuk, V., Petrenko, A., Singh, R., Liu, J., Makoviychuk, D., Van Wyk, K., Zhurkevich, A., Sundaralingam, B., et al.: Dextreme: Transfer of agile in-hand manipulation from simulation to reality. In: *2023 IEEE*

- International Conference on Robotics and Automation (ICRA). pp. 5977–5984. IEEE (2023)
18. Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–9 (2023)
 19. He, J., Li, D., Yu, X., Qi, Z., Zhang, W., Chen, J., Zhang, Z., Zhang, Z., Yi, L., Wang, H.: Dexvlg: Dexterous vision-language-grasp model at scale. arXiv preprint arXiv:2507.02747 (2025)
 20. He, T., Luo, Z., He, X., Xiao, W., Zhang, C., Zhang, W., Kitani, K., Liu, C., Shi, G.: Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. arXiv preprint arXiv:2406.08858 (2024)
 21. Hendrycks, D., Gimpel, K.: Gaussian error linear units (GELUs) (Jun 2016)
 22. Huang, B., Wang, Y., Yang, X., Luo, Y., Li, Y.: 3d-vitac: Learning fine-grained manipulation with visuo-tactile sensing. arXiv preprint arXiv:2410.24091 (2024)
 23. Kareer, S., Patel, D., Punamiya, R., Mathur, P., Cheng, S., Wang, C., Hoffman, J., Xu, D.: Egomimic: Scaling imitation learning via egocentric video. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 13226–13233. IEEE (2025)
 24. Kim, J., Kim, J., Na, J., Joo, H.: Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 1816–1828 (2025)
 25. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10138–10148 (2021)
 26. Lee, J., Park, E., Cho, M.: Dexter: Language-driven dexterous grasp generation with embodied reasoning. arXiv preprint arXiv:2601.16046 (2026)
 27. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
 28. Li, J., Zhu, Y., Xie, Y., Jiang, Z., Seo, M., Pavlakos, G., Zhu, Y.: Okami: Teaching humanoid robots manipulation skills through single video imitation. arXiv preprint arXiv:2410.11792 (2024)
 29. Li, K., Li, P., Liu, T., Li, Y., Huang, S.: Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6991–7003 (2025)
 30. Li, K., Wang, J., Yang, L., Lu, C., Dai, B.: Semgrasp: Semantic grasp generation via language aligned discretization. In: European Conference on Computer Vision. pp. 109–127. Springer (2024)
 31. Liao, Q., Truong, T.E., Huang, X., Gao, Y., Tevet, G., Sreenath, K., Liu, C.K.: Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. arXiv preprint arXiv:2508.08241 (2025)
 32. Lu, J., Kang, H., Li, H., Liu, B., Yang, Y., Huang, Q., Hua, G.: Ugg: Unified generative grasping. In: European Conference on Computer Vision. pp. 414–433. Springer (2024)
 33. Luo, Z., Cao, J., Christen, S., Winkler, A., Kitani, K., Xu, W.: Omnigrasp: Grasping diverse objects with simulated humanoids. Advances in Neural Information Processing Systems **37**, 2161–2184 (2024)
 34. Luo, Z., Cao, J., Kitani, K., Xu, W., et al.: Perpetual humanoid control for real-time simulated avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10895–10904 (2023)

35. Luo, Z., Cao, J., Merel, J., Winkler, A., Huang, J., Kitani, K., Xu, W.: Universal humanoid motion representations for physics-based control. *arXiv preprint arXiv:2310.04582* (2023)
36. Luo, Z., Yuan, Y., Wang, T., Li, C., Chen, S., Castaneda, F., Cao, Z.A., Li, J., Minor, D., Ben, Q., et al.: Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820* (2025)
37. Mahmood, N., Ghorbani, N., Troje, N.F., Pons-Moll, G., Black, M.J.: Amass: Archive of motion capture as surface shapes. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5442–5451 (2019)
38. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., State, G.: Isaac gym: High performance gpu-based physics simulation for robot learning. In: *NeurIPS Datasets and Benchmarks* (2021)
39. Mandi, Z., Hou, Y., Fox, D., Narang, Y., Mandlekar, A., Song, S.: Dexmachina: Functional retargeting for bimanual dexterous manipulation. *arXiv preprint arXiv:2505.24853* (2025)
40. Meng, H., Jin, S., Liu, W., Qian, C., Lin, M., Ouyang, W., Luo, P.: 3d interacting hand pose estimation by hand de-occlusion and removal. In: *European Conference on Computer Vision*. pp. 380–397. Springer (2022)
41. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: *European Conference on Computer Vision*. pp. 548–564. Springer (2020)
42. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12999–13008 (2023)
43. Pan, C., Wang, C., Qi, H., Liu, Z., Bharadhwaj, H., Sharma, A., Wu, T., Shi, G., Malik, J., Hogan, F.: Spider: Scalable physics-informed dexterous retargeting. *arXiv preprint arXiv:2511.09484* (2025)
44. Pan, L., Yang, Z., Dou, Z., Wang, W., Huang, B., Dai, B., Komura, T., Wang, J.: Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5379–5391 (2025)
45. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 10975–10985 (2019)
46. Peng, X.B., Abbeel, P., Levine, S., Van de Panne, M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)* **37**(4), 1–14 (2018)
47. Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)* **41**(4), 1–17 (2022)
48. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)* **40**(4), 1–20 (2021)
49. Qi, H., Yi, B., Suresh, S., Lambeta, M., Ma, Y., Calandra, R., Malik, J.: General in-hand object rotation with vision and touch. In: *Conference on Robot Learning*. pp. 2549–2564. PMLR (2023)

50. Qin, Y., Su, H., Wang, X.: From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation. *IEEE Robotics and Automation Letters* **7**(4), 10873–10881 (2022)
51. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017)
52. Si, Z., Zhang, G., Ben, Q., Romero, B., Xian, Z., Liu, C., Gan, C.: Diff tactile: A physics-based differentiable tactile simulator for contact-rich robotic manipulation. *arXiv preprint arXiv:2403.08716* (2024)
53. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: *European conference on computer vision*. pp. 581–600. Springer (2020)
54. Tessler, C., Jiang, Y., Coumans, E., Luo, Z., Chechik, G., Peng, X.B.: Masked-manipulator: Versatile whole-body control for loco-manipulation. *arXiv preprint arXiv:2505.19086* (2025)
55. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: Clod: Closing the loop between simulation and diffusion for multi-task character control. *arXiv preprint arXiv:2410.03441* (2024)
56. Turpin, D., Wang, L., Heiden, E., Chen, Y.C., Macklin, M., Tsogkas, S., Dickinson, S., Garg, A.: Grasp’d: Differentiable contact-rich grasp synthesis for multi-fingered hands. In: *European Conference on Computer Vision*. pp. 201–221. Springer (2022)
57. Wang, R., Zhang, J., Chen, J., Xu, Y., Li, P., Liu, T., Wang, H.: Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In: *IEEE International Conference on Robotics and Automation (ICRA)*. pp. 11359–11366 (2023)
58. Wang, Y., Zhao, Q., Yu, R., Tsui, H.W., Zeng, A., Lin, J., Luo, Z., Yu, J., Li, X., Chen, Q., et al.: Skillmimic: Learning basketball interaction skills from demonstrations. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 17540–17549 (2025)
59. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. *Advances in Neural Information Processing Systems* **37**, 46881–46907 (2024)
60. Wei, Y.L., Lin, M., Lin, Y., Jiang, J.J., Wu, X.M., Zeng, L.A., Zheng, W.S.: Afford-dexgrasp: Open-set language-guided dexterous grasp with generalizable-instructive affordance. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11818–11828 (2025)
61. Weng, Z., Lu, H., Kragic, D., Lundell, J.: Dexdiffuser: Generating dexterous grasps with diffusion models. *IEEE Robotics and Automation Letters* **9**(12), 11834–11840 (2024)
62. Won, J., Gopinath, D., Hodgins, J.: A scalable approach to control diverse behaviors for physically simulated characters. *ACM Trans. Graph.* **39**(4) (2020), <https://doi.org/10.1145/3386569.3392381>
63. Wu, Z., Li, J., Xu, P., Liu, C.K.: Human-object interaction from human-level instructions. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 11176–11186 (2025)
64. Xu, G.H., Wei, Y.L., Zheng, D., Wu, X.M., Zheng, W.S.: Dexterous grasp transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 17933–17942 (2024)
65. Xu, M., Zhang, H., Hou, Y., Xu, Z., Fan, L., Veloso, M., Song, S.: Dexumi: Using human hand as the universal manipulation interface for dexterous manipulation. *arXiv preprint arXiv:2505.21864* (2025)

66. Xu, S., Ling, H.Y., Wang, Y.X., Gui, L.Y.: Intermimic: Towards universal whole-body control for physics-based human-object interactions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12266–12277 (2025)
67. Xu, Y., Wan, W., Zhang, J., Liu, H., Shan, Z., Shen, H., Wang, R., Geng, H., Weng, Y., Chen, J., et al.: Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4737–4746 (2023)
68. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: 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)
69. Yin, K., Zeng, W., Fan, K., Dai, M., Wang, Z., Zhang, Q., Tian, Z., Wang, J., Pang, J., Zhang, W.: Unitracker: Learning universal whole-body motion tracker for humanoid robots. arXiv preprint arXiv:2507.07356 (2025)
70. Yu, K., Han, Y., Wang, Q., Saxena, V., Xu, D., Zhao, Y.: Mimictouch: Leveraging multi-modal human tactile demonstrations for contact-rich manipulation. arXiv preprint arXiv:2310.16917 (2023)
71. Yu, R., Wang, Y., Zhao, Q., Tsui, H.W., Wang, J., Tan, P., Chen, Q.: Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
72. Yuan, Z., Wei, T., Gu, L., Hua, P., Liang, T., Chen, Y., Xu, H.: Hermes: Human-to-robot embodied learning from multi-source motion data for mobile dexterous manipulation. arXiv preprint arXiv:2508.20085 (2025)
73. Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., Lu, C.: Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 445–456 (2024)
74. Zhang, H., Christen, S., Fan, Z., Hilliges, O., Song, J.: Grasppl: Generating grasping motions for diverse objects at scale. In: European Conference on Computer Vision. pp. 386–403. Springer (2024)
75. Zhang, Z., Guo, J., Chen, C., Wang, J., Lin, C., Lian, Y., Xue, H., Wang, Z., Liu, M., Lyu, J., et al.: Track any motions under any disturbances. arXiv preprint arXiv:2509.13833 (2025)
76. Zheng, R., Niu, D., Xie, Y., Wang, J., Xu, M., Jiang, Y., Castañeda, F., Hu, F., Tan, Y.L., Fu, L., et al.: Egoscale: Scaling dexterous manipulation with diverse egocentric human data. arXiv preprint arXiv:2602.16710 (2026)
77. Zhong, Y., Huang, X., Li, R., Zhang, C., Chen, Z., Guan, T., Zeng, F., Lui, K.N., Ye, Y., Liang, Y., et al.: Dexgraspvla: A vision-language-action framework towards general dexterous grasping. arXiv preprint arXiv:2502.20900 (2025)
78. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 813–822 (2019)