PiHand: Latent Hand–Object Interaction Priors for Physics-Based Policy Learning

Anonymous authors
Paper under double-blind review

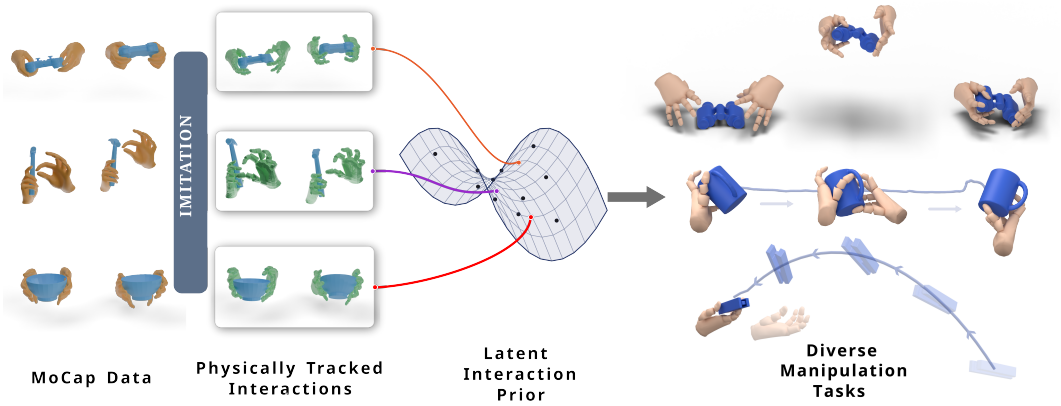


Figure 1: We propose PiHand, a method to generate physically plausible hand-object interaction motions by learning physically grounded interaction priors from mocap references. Our single interaction prior transfers to diverse dexterous manipulation tasks.

Abstract

Synthesizing physically grounded dexterous hand–object interactions is important for VR/AR, gaming, and robotics, yet remains challenging because subtle contact dynamics, diverse object geometries, and target trajectories must be jointly respected while maintaining natural grasps. Existing methods either track mocap trajectories with limited transfer to novel objects, or rely on hand-only generative priors that fail to capture coupled hand–object motion, sacrificing contact fidelity and sample efficiency. To tackle these challenges, we introduce **PiHand**, a **P**hysically grounded **I**nteraction-aware **Hand** motion synthesis framework. PiHand first trains expert policies that track mocap demonstrations in simulation, distills them into a unified hand–object interaction prior using a goal-conditioned VAE, and finally solve diverse dexterous manipulation tasks through bounded residual exploration in its latent space. With only a single reward and no complex curriculum training, this yields general tabletop manipulation policies that generalize to unseen objects and trajectories on GRAB and DexYCB. The same prior transfers to object catching and bimanual handover, two tasks that are rarely seen in its training data. Across all tasks, PiHand consistently outperforms baselines in all metrics. We will release the processed data and model checkpoints.

1 Introduction

Generating realistic hand-object interactions is a foundational challenge in computer vision and graphics, with implications for VR/AR, gaming and robotics. Several kinematic approaches (Cha et al., 2024; Li et al., 2025b; Zhang et al., 2025b) have made significant strides in generating such interactions from visual or textual inputs, yet they frequently suffer from noticeable physical implausibilities which make them unfit for deployment in simulated environments or on real-world robots. To bridge this gap, the robotics and physics-based animation communities have focused on learning generative control policies to synthesize the complex motions required to stably grasp and

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

manipulate objects based on their geometry and target trajectory inside simulation. Yet this remains challenging because contact dynamics are subtle and discrete, objects vary widely in geometry and affordances, and the hand must continuously adapt its grasp and contact configuration to remain both natural-looking and physically correct.

Physics-based tracking controllers imitate mocap sequences (Li et al., 2025a; Liu et al., 2025a) with strong realism and sim-to-real transfer, but remain reconstructive: their reliance on dense references prevents synthesis or adaptation to new objects without corresponding mocap. Conversely, generative RL methods such as GraspXL (Zhang et al., 2025a) generalize to unseen objects. But such dexterous-hand exploration in action space requires engineered rewards and curricula, remains sample-inefficient, and often yields unnatural motion. PULSE (Luo et al., 2024b) learns reusable latent humanoid control spaces, which Omnigrasp (Luo et al., 2024a) applies to object-centric tasks. Yet these human-only priors do not model object interactions, forcing policies to discover contacts in a latent space unstructured for them and limiting scalability to complex, dynamic tasks.

In this work, we demonstrate that an *interaction prior* consistently outperforms both action-space exploration and hand-only priors in sample efficiency, spatial accuracy and tracking robustness. We propose **PiHand**, a **P**hysically grounded **I**nteraction-aware **Hand** motion synthesis framework that efficiently generates high-quality hand-object interactions, generalizes to new objects and transfers to unseen dexterous tasks. Instead of relying on brittle reward engineering or "blind" motion priors, our core insight is to embed the interaction itself into the foundation of the latent space. We argue that such an "interaction-first" approach is essential to capture the subtle, contact-rich patterns of human grasping and to construct a well-behaved latent manifold that enables efficient synthesis across novel objects, trajectories and tasks. Specifically, we propose a three-stage pipeline: we first train specialized expert policies in a physics simulator to track hand–object interactions from mocap data, then distill these experts into an object- and contact-aware latent interaction prior, and finally solve new tasks by residual RL exploration within this latent space.

Experiments show that our imitation experts improve success rate (SR) over the state-of-the-art tracking method from 69.6% to 93.3%, providing accurate data for distillation. Using only a simple object-tracking reward, our interaction prior enables tabletop policies to manipulate unseen objects along unseen trajectories with 14.0 and 16.9 percentage points higher SR on GRAB (Taheri et al., 2020) and DexYCB (Chao et al., 2021), respectively. The same prior transfers to object catching absent from its training data and sparsely observed bimanual handover, achieving $2.0\times$ and $4.3\times$ the strongest-baseline SR, respectively.

In summary, our main contributions are:

- **Data-Driven Interaction Prior for Efficient RL:** We introduce a data-driven *hand-object interaction prior* that allows efficient RL policy learning for dexterous manipulation, bridging mocap-based human priors with physics-based reinforcement learning.
- **High-Fidelity Interaction Experts:** We learn physically accurate interaction experts to track reference mocap demonstrations for the GRAB dataset, providing a valuable resource for the community and a foundation for learning from human demonstrations.
- **Generalizable and Reusable Interaction Synthesis:** We show that our learned single interaction prior supports unseen objects and trajectories on GRAB and DexYCB, transfers to rare bimanual handovers and unseen object-catching tasks, and consistently outperforms baselines in success rate and spatial controllability.

2 Related Work

Human and interaction motion imitation. Physics-based motion imitation typically learns reference-conditioned tracking controllers from human motion. Starting from example-guided RL (DeepMimic (Peng et al., 2018)) and adversarial motion priors/skill embeddings (AMP (Peng et al., 2021), ASE (Peng et al., 2022)), recent general motion tracking systems scale to diverse whole-body behaviors and hardware transfer, e.g., PHC (Luo et al., 2023), HOVER (He et al., 2025), ExBody2 (Ji et al., 2024), GMT (Chen et al., 2025), KungfuBot (Xie et al., 2025), and camera/teleop-driven imitators such as H2O (He et al., 2024) and TWIST (Ze et al., 2025). More recently, VideoMimic (Allshire et al., 2025) imitates human motion from video and transfers it to a real robot.

Extending imitation to human–object interaction (HOI) requires contact-aware objectives and diverse interaction data: PhysHOI (Wang et al., 2023) introduces contact-graph rewards; SkillMimic (Wang et al., 2025b) and SkillMimic-V2 (Yu et al., 2025) learn reusable interaction skills from demonstrations and InterMimic (Xu et al., 2025) scales to varied objects via teacher–student distillation. Recent loco-manipulation frameworks further build on these trackers with residual refinement or generative priors (MaskedManipulator (Tessler et al., 2025), ResMimic (Zhao et al., 2025), DreamControl (Kalaria et al., 2025), VisualMimic (Yin et al., 2025), CoorDex (Li et al., 2026)).

For hand–object interaction, DexTrack (Liu et al., 2025a), ManipTrans (Li et al., 2025a), AdaDex-Track (Adalibieke et al., 2026), HOT (Wang et al., 2025a), and DexMachina (Mandi et al., 2025) learn dexterous tracking control from human references, while DexMimicGen (Jiang et al., 2025) and video-based pipelines (DexMV (Qin et al., 2021), ViViDex (Chen et al., 2024)) expand demonstration coverage; classic demo-augmented RL (DAPG (Rajeswaran et al., 2017)) underpins many of these systems. In contrast to predominantly reference-following controllers (or guidance-based synthesis), our method additionally supports generation of novel hand-object motions beyond tracking.

Interaction motion generation. Physics-based approaches often learn task-specific policies for basketball (Wang et al., 2025b; Yu et al., 2025), tennis (Zhang et al., 2023), bulky-object manipulation (Zheng et al., 2025), or in-hand rotation (Liu et al., 2025b). Such pipelines require careful reward shaping and curricula and do not readily scale across tasks. To improve reuse and data efficiency, prior work incorporates human-motion representations, ranging from imitation-RL controllers such as DeepMimic (Peng et al., 2018), AMP (Peng et al., 2021), and PHC (Luo et al., 2023) to reusable skill priors such as ASE (Peng et al., 2022). PULSE (Luo et al., 2024b) further distills motion-tracking behavior into a conditional latent representation for downstream latent-space RL. Its overall recipe is the closest precursor to PiHand. We follow this recipe but replace PULSE's human-only whole-body representation with a dexterous hand–object prior that jointly models hand state, object geometry and trajectory, and contact dynamics.

Building on these priors, recent works move toward modelling object manipulations: Omnigrasp leverages pretrained humanoid motion representations to learn grasp-and-carry across more than 1,000 objects (Luo et al., 2024a), GraspXL scales goal-conditioned grasp-motion synthesis to massive unseen object sets (Zhang et al., 2025a), and MaskedMimic reframes physics-based control as masked motion inpainting conditioned on sparse cues (e.g., keyframes/text/object goals) (Tessler et al., 2024). TokenHSI further unifies diverse physical human–scene skills via task tokenization in a single transformer policy (Pan et al., 2025), but still largely relies on human-only priors. Only a few works directly learn interaction priors: concurrent work InterPrior (Xu et al., 2026) distills full-reference HOI experts into a goal-conditioned variational controller and RL-finetunes for generalization, complementary to kinematic HOI priors that explicitly model human–object coupling (DecHOI (Jung et al., 2025), CoopDiff (Lin et al., 2025)); MaskedManipulator similarly distills a tracking expert into a goal-conditioned generative loco-manipulation policy over coupled body–object objectives (Tessler et al., 2025). Orthogonal to these, some kinematic methods (Li et al., 2023; 2024; Zhang et al., 2022; 2024; Zhou et al., 2024; Taheri et al., 2022; Wu et al., 2025; Peng et al., 2023) learn interaction priors by design but their physical plausibility is limited without simulation. Overall, many systems still rely on human-only priors or task-specific rewards rather than a learned interaction prior, motivating our interaction-first approach for more efficient learning and accurate manipulation.

3 METHOD

Our method synthesizes physically plausible hand–object interactions conditioned only on a target object trajectory. The key idea is to construct a latent space that embeds data-driven priors of physically simulated hand-object interaction motions. As illustrated in Figure 2, our pipeline consists of three stages. In Stage I we learn a set of per-object expert policies that convert noisy mocap demonstrations into physically grounded hand-object interaction motions. In Stage II, we distill the expert's behaviors into a latent interaction prior. Finally, in Stage III we explore this latent space, allowing the model to efficiently generate new interactions and solve novel tasks.

Preliminaries. Our simulation environment contains two PD-actuated hands, an unactuated rigid object, and a static table. Our **hand state** $\mathbf{q} = (\mathbf{q}^L, \mathbf{q}^R)$ consists of the left hand $\mathbf{q}^L$ and the right hand $\mathbf{q}^R$ where each hand $\mathbf{q}^h = (\mathbf{P}^h_w, \boldsymbol{\theta}^h), h \in \{L, R\}$ is represented by the wrist pose $\mathbf{P}^h_w \in SE(3)$ and internal finger joint angles $\boldsymbol{\theta}^h \in \mathbb{R}^{45}$. Each hand has a total of $n_D = 51$ DoF, $n_B = 16$ collision

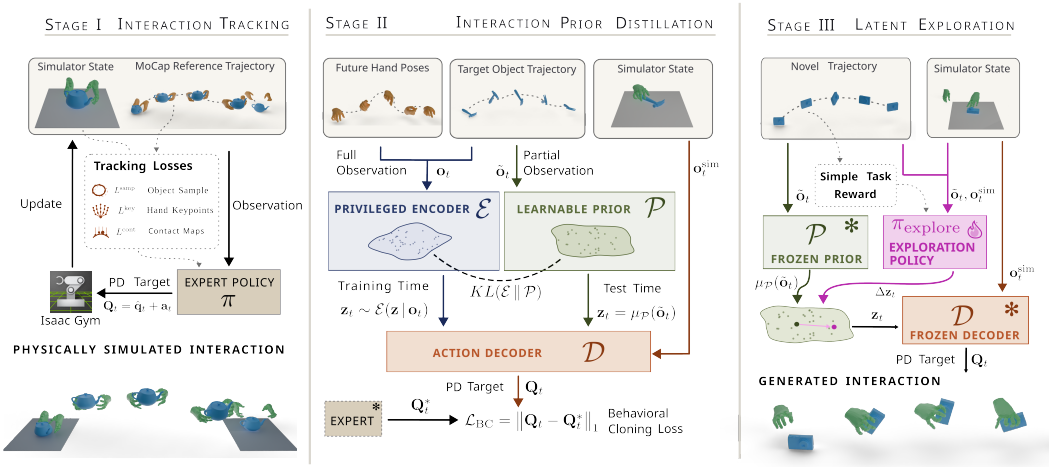


Figure 2: **Method Overview**. In Stage I we train expert policies that can track mocap hand-object interactions in a physically simulated environment. In Stage II we then distill the expert knowledge into a VAE-like student network via behavior cloning. The VAE embeds hand-object interactions into a latent space which we leverage in Stage III for efficient sampling and exploration for object trajectory guided motion generation.

bodies, and we additionally use $n_K = 21$ kinematic keypoints per hand, including the wrist, internal joints, and fingertips, to compute rewards on hand pose and contact. The passive object is represented by its pose $\mathbf{P}_o \in SE(3)$ and linear and rotational velocities $\dot{\mathbf{P}}_o$, and evolves only under gravity and contact forces. We use $\hat{\cdot}$ to denote reference quantities from mocap throughout this section. We treat $\mathbf{P} = (\mathbf{R}, \mathbf{t}) \in SE(3)$ as acting on points by $\mathbf{Px} = \mathbf{Rx} + \mathbf{t}$.

3.1 Stage I: Physics-Grounded Hand–Object Interaction Tracking

Mocap demonstrations record kinematics but not dynamics: contacts are frequently inconsistent and forces are absent altogether. In Stage I we aim to produce physically valid interaction data by training expert tracking policies via PPO (Schulman et al., 2017) to imitate mocap demonstrations in a simulator. The resulting experts are then used in Stage II to learn a latent interaction prior for motion generation.

Because the mocap reference is already a strong prior on the desired motion, the policy need only correct it rather than synthesize control from scratch. We therefore parameterize the PD target at time t as the reference-anchored pose $\mathbf{Q}_t = \hat{\mathbf{q}}_t + \mathbf{a}_t$, where $\mathbf{a}_t \in \mathbb{R}^{2n_D}$ is a learned residual correction in pose space. Predicting $\mathbf{a}_t$ directly, however, allows the policy to make large frame-to-frame jumps in the PD target, which produces jittery control and unstable contacts. Instead, π outputs a per-step *increment* that we integrate, $\mathbf{a}_t = \mathbf{a}_{t-1} + \pi(\mathbf{o}_t)$, with $\mathbf{a}_0 = \mathbf{0}$, so that temporal smoothness is built into the action parameterization. We further constrain $\mathbf{a}_t$ to per-joint deviation bounds around $\hat{\mathbf{q}}_t$ and clamp the PD target to natural joint limits.

Input observation $\mathbf{o}_t$ decomposes into three parts, $\mathbf{o}_t = (\mathbf{o}_t^{\text{sim}}, \mathbf{o}_t^{\text{ref}}, \mathbf{a}_{t-1})$, where $\mathbf{o}_t^{\text{sim}}$ is the current simulator state, $\mathbf{o}_t^{\text{ref}}$ is a future reference goal, and $\mathbf{a}_{t-1}$ is the previously predicted residual correction term. To capture the hand-object interaction, the simulator state $\mathbf{o}_t^{\text{sim}}$ consists of hand poses $\mathbf{q}$, hand keypoints $\mathbf{K} \in \mathbb{R}^{2n_K \times 3}$, object states $\mathbf{P}_o$, contact indicators $\mathbf{c} \in \{0,1\}^{2n_B}$, and explicit displacement vectors $\mathbf{D} \in \mathbb{R}^{2n_K \times 3}$ from each hand keypoint to the nearest sample on the object surface, computed against a fixed canonical set $\{\mathbf{s}_k\}_{k=1}^{K}$ rigidly transformed into the world frame by the current object pose.

We construct the reference goal observation $\mathbf{o}_t^{\text{ref}}$ using the future mocap reference in a look-ahead horizon $\tau \in \mathcal{T}$ (sampled densely near t, sparsely out to 2 s). The goal $\mathbf{o}_t^{\text{ref}}$ stacks future mocap targets: reference object pose $\hat{\mathbf{P}}_{o,\tau}$, hand poses $\hat{\mathbf{q}}_\tau$, and keypoints $\hat{\mathbf{K}}_\tau$ and their corresponding residuals to the current state $\Delta\hat{\mathbf{P}}_{o,\tau}, \Delta\hat{\mathbf{q}}_\tau, \Delta\hat{\mathbf{K}}_\tau$ together with the reference binary contact map $\hat{\mathbf{c}}_\tau$ into the observation vector. We use the 6D representation (Zhou et al., 2019) for all rotations.

Contact-aware tracking reward. We supervise the policy with a per-step reward that combines three tracking losses through exponential shaping: $r_t = \sum_{l \in \{L^{\text{samp}}, L^{\text{cont}}, L^{\text{key}}\}} \exp(-\lambda_l \, l)$ where the object surface-sample loss L^{samp} tracks the rigid object pose, the contact loss L^{cont} tracks the reference contact map, and the keypoint loss L^{key} tracks the hand keypoint trajectory. We detail each loss next.

The *object surface-sample loss* $L^{\text{samp}} = \frac{1}{K} \sum_{k=1}^{K} \left\| \mathbf{P}_o \mathbf{s}_k - \hat{\mathbf{P}}_o \mathbf{s}_k \right\|_2$ captures object translation and rotation by tracking a fixed set of canonical object surface points $\{\mathbf{s}_k\}_{k=1}^{K}$ in the world frame.

The *contact loss* $L^{\text{cont}} = \frac{1}{2n_B} \sum_{j=1}^{2n_B} \max\big(\hat{\mathbf{c}}_j - \mathbf{c}_j, \, 0\big)$ enforces reference contacts as a one-sided penalty on missed contacts. We deliberately do not penalize spurious contacts symmetrically: mocap contact maps are geometrically inferred and noisy, and a symmetric penalty would push the policy off physically plausible grasps.

For keypoint tracking, we combine two terms. The *pose loss* $\ell_i^{\text{pose}} = \|\mathbf{P}_w^{-1}\mathbf{K}_i - \hat{\mathbf{P}}_w^{-1}\hat{\mathbf{K}}_i\|_2$ aims to minimize the distance between simulated and reference finger keypoints, expressed in their respective wrist frame, pushing the local pose of the hand to be consistent, independent of wrist deviations. The *contact position loss* $\ell_i^{\text{copo}} = \|\mathbf{P}_o^{-1}\mathbf{K}_i - \hat{\mathbf{P}}_o^{-1}\hat{\mathbf{K}}_i\|_2$, expressed in object frames, drives the simulated fingers to meet the object at the same surface locations as the reference. These terms are combined into the final *keypoint loss* $L^{\text{key}} = \frac{1}{2n_K} \sum_{i=1}^{2n_K} \big(w_i \, \ell_i^{\text{copo}} + (1 - w_i) \, \ell_i^{\text{pose}}\big)$, where the weighting w_i shifts towards ℓ_i^{copo} as the keypoint approaches a reference contact event. Details on the computation of w_i are given in Appendix B.

Per-object experts and hard-mining. Rather than training one expert per sequence, which is expensive, we train two generalist experts for each object where each expert handles either single or bimanual interactions respectively. We prune unrecoverable rollouts via early termination (Peng et al., 2018): episodes end on hand self-penetration or on excessive deviation from the reference measured by L^{samp}, with the deviation threshold annealed as a curriculum. We further apply a hard-mining sampling strategy that biases the expert's sequence pool toward low recent tracking-completion rates. See Appendix B for details. With this training curriculum, our per-sequence experts outperform ManipTrans (Li et al., 2025a), and our per-object experts maintain a high success rate on the full dataset (Table 1).

3.2 Stage II: Interaction Prior Distillation

The experts trained in Stage I are privileged policies that can physically reproduce mocap interactions, but are not generative themselves, as they require full reference hand motions that are unavailable at generation time. The goal of Stage II is to compress the entire expert pool into a single goal-conditioned generative model whose *learnable prior* and *decoder* together act as an *interaction-aware* prior over physically simulated hand-object motions. This prior–decoder pair is then used in Stage III as the latent backbone for RL exploration. Unlike prior motion-generation works that model human kinematics (Luo et al., 2024a; Tessler et al., 2024) without the object, our prior captures the coupled hand–object dynamics distilled from physically valid expert behavior.

Object-trajectory-conditioned generative model. We instantiate the model as a conditional VAE (Kingma & Welling, 2013; Sohn et al., 2015) with three components (see middle of Figure 2) sharing a Gaussian latent space. The *learnable prior* $\mathcal{P}(\mathbf{z}_t \mid \tilde{\mathbf{o}}_t)$ is conditioned on the *partial observation* $\tilde{\mathbf{o}}_t$, which contains only the future object-trajectory portion of the reference goal. The privileged *encoder* $\mathcal{E}(\mathbf{z}_t \mid \mathbf{o}_t)$ is conditioned on the *privileged observation* $\mathbf{o}_t$, which extends $\tilde{\mathbf{o}}_t$ with the reference hand trajectory features of the reference goal. The deterministic *decoder* $\mathcal{D}(\mathbf{o}_t^{\text{sim}}, \mathbf{z}_t)$ maps the current simulation state $\mathbf{o}_t^{\text{sim}}$ and a sampled latent to a per-step update on the simulated hand pose, which we apply as a residual on the current pose to form the PD target $\mathbf{Q}_t = \mathbf{q}_{t-1} + \mathcal{D}(\mathbf{o}_t^{\text{sim}}, \mathbf{z}_t)$. At training time, latents are drawn from the encoder, $\mathbf{z}_t \sim \mathcal{E}(\cdot \mid \mathbf{o}_t)$; at test time the encoder is discarded and we use the prior's mean $\mu_{\mathcal{P}}(\tilde{\mathbf{o}}_t)$ so generation never depends on reference hand poses.

Training losses. We train our model with behavior-cloning loss $\mathcal{L}_{\text{BC}}$ and weighted Kullback-Leibler (KL) divergence loss $\mathcal{D}_{\text{KL}}$: $\mathcal{L}_{\text{distill}} = \mathcal{L}_{\text{BC}} + \beta \, \mathcal{D}_{\text{KL}}$. Here $\mathcal{L}_{\text{BC}}$ matches the decoder output to the expert's PD target on every visited state. Rollouts follow DAgger (Ross et al., 2011): we use the decoder to run simulation in the environment, while the experts are queried for a supervision action. At every time step, we match the student's PD target $\mathbf{Q}_t$ to the expert's PD target $\mathbf{Q}_t^*$ produced by

the Sec. 3.1 policy using L_1 loss: $\mathcal{L}_{\text{BC}} = \|\mathbf{Q}_t - \mathbf{Q}_t^*\|_1$. Matching at the PD-target level absorbs the reference-pose anchoring built into the expert's residual formulation, so the decoder learns natural per-step updates rather than reconstructing the anchoring offset. The KL term $\mathcal{D}_{\text{KL}} = \text{KL}\big(\mathcal{E}(\mathbf{z}_t \mid \mathbf{o}_t) \,\|\, \mathcal{P}(\mathbf{z}_t \mid \tilde{\mathbf{o}}_t)\big)$ aligns the partial-observation prior with the privileged-observation encoder, so that latents sampled from $\mathcal{P}$ at test time stay on the manifold the decoder was trained against.

3.3 Stage III: Interaction Generation via Bounded Latent Exploration

Stage II distillation yields a compact, interaction-aware latent space whose samples decode to physically plausible actions. To generalize beyond the limited training objects and interactions to novel objects, trajectories, and tasks, we explore this space with RL. The learned interaction prior improves sample efficiency and suppresses unnatural motion, providing a general approach to different manipulation tasks (Sec. 4).

Bounded latent exploration setup. Independent of the concrete task, we train an exploration policy π_{explore} via PPO that consumes the partial observation $\tilde{\mathbf{o}}_t$ and the current simulator state $\mathbf{o}_t^{\text{sim}}$, and outputs a residual latent $\Delta\mathbf{z}_t$ on top of the frozen prior's mean. The action is then produced by the frozen decoder $\mathbf{a}_t = \mathcal{D}\big(\mathbf{o}_t^{\text{sim}},\, \mu_{\mathcal{P}}(\tilde{\mathbf{o}}_t) + \Delta\mathbf{z}_t\big)$. To keep the latent close to the original space, we bound the magnitude of $\Delta\mathbf{z}_t$ by a constant B (similar to Yu et al., 2026; Zhang et al., 2026) via a smooth squash $\Delta\mathbf{z} \leftarrow B\tanh\big(\frac{\|\Delta\mathbf{z}\|}{B}\big)\frac{\Delta\mathbf{z}}{\|\Delta\mathbf{z}\|}$. We show in Sec. 4.4 that this bound constrains the exploration and maintains synthesis diversity.

4 Experiments

We compare our Stage I experts with a state-of-the-art mocap tracker and our interaction-prior generator with baselines on tabletop manipulation, object catching, and bimanual handover, consistently improving success rates. See the supplementary video for qualitative results.

Datasets and Implementation Details. We use GRAB (Taheri et al., 2020) as our dataset for training the interaction prior, but also rely on DexYCB (Chao et al., 2021) as a secondary source to show the cross-dataset generalization capabilities of our GRAB-trained interaction prior. All experiments are done within IsaacGym (Makoviychuk et al., 2021) with MANO-consistent robotic hands. Further details on data split, preprocessing and simulator can be found in Appendix C.

Evaluation protocol and metrics. Unless otherwise stated, we evaluate each method using 100 simulated rollouts per sequence to account for simulator stochasticity. A rollout is terminated when the object deviates from the reference trajectory beyond a task-dependent distance threshold. To measure temporal robustness, we report *tracking completion (TC in $\%$)*, defined as the fraction of the reference trajectory completed before termination, and *success rate (SR in $\%$)*, defined as the fraction of rollouts that complete the entire sequence. For spatial precision, we report average *object translation error (E_t in cm)*, the *object rotation error (E_r in $^\circ$)*, and, for the Stage I imitation task where reference motions are available, the *mean per joint position error (MPJPE in cm)* of the hands.

4.1 Hand–Object Interaction Imitation

We compare our Stage I experts with ManipTrans (Li et al., 2025a), a state-of-the-art per-sequence imitation method. Since ManipTrans does not support bimanual manipulation of single objects, we compare with it on GRAB50 which is a random subset of 50 single-handed GRAB (Taheri et al., 2020) sequences. As Table 1 shows, our experts trained per-sequence outperform ManipTrans trained on an identical setup on all metrics. We also train two experts per object, one for single-handed and one for bimanual interactions, on the full GRAB dataset. While the success rate slightly degrades, it is much more scalable than training one expert per sequence as done in ManipTrans and our overall success rate is still better than per-sequence ManipTrans experts. This shows that our method is a scalable data engine for learning physical interaction priors from human demonstrations.

4.2 Tabletop Manipulation

Our main application of the interaction prior is a task we call *Tabletop Manipulation*: given only a target object trajectory, a single policy must synthesize a physically plausible interaction for objects



Figure 3: **Tabletop manipulation synthesis comparison** on unseen objects from GRAB (top row) and DexYCB (bottom row) evaluation set. Grayed out frames indicate early termination due to fatally wrong grasps. Action space exploration generates unnatural grasping and hand-only prior often fails to grasp with object reference while ours achieves both realism and high success rate.

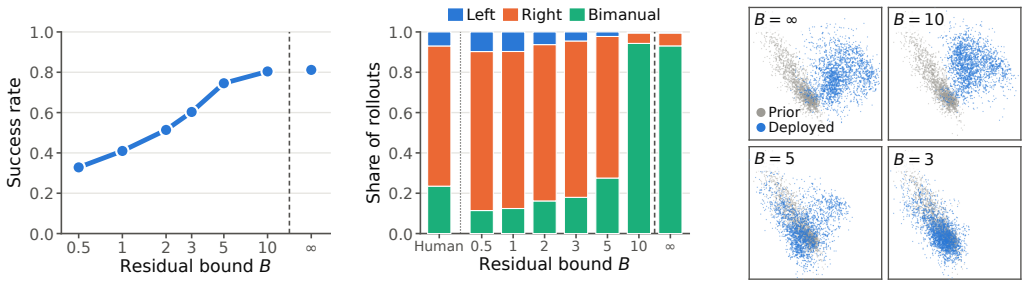
Table 1: **Physical imitation of mocap reference motions** on GRAB50 and the full GRAB dataset. Our method significantly improves the tracking success rate than ManipTrans (Li et al., 2025a).

Method	Data	Granularity	SR ↑	TC ↑	E_t (cm)↓	E_r (°)↓	MPJPE (cm) ↓
ManipTrans	GRAB50	Per-sequence	69.6	77.6	0.43	7.3	1.97
Ours	GRAB50	Per-sequence	**93.3**	**95.8**	**0.28**	**4.9**	**1.01**
	GRAB Full	Per-object	84.8	92.2	1.04	15.4	2.08

and trajectories it has never seen. To this end, we train the Stage III residual policy π_{explore} over a broad distribution of initial states and target trajectories.

Setup. Each training episode initializes the hands and the object from two *different* training sequences, so the policy starts from hand–object configurations that neither the experts nor the prior have encountered. The target trajectory is, with equal probability, an original training trajectory or a procedurally generated lift-and-move trajectory. The policy only ever interacts with training objects, whereas at evaluation both objects and trajectories are held out, making the policy zero-shot with respect to object geometry and motion. We train one Stage III policy per dataset, on GRAB and on DexYCB, each on the respective training objects and on top of the same GRAB-trained prior. Since generation is considerably harder than imitation, we relax the termination threshold for SR and TC to 10 cm. However, our advantage persists and even grows under stricter thresholds (Appendix F.2).

Baselines. We implement two alternatives within our framework, so that simulator, embodiment and physics are identical across methods (see Appendix D). **Action Space Exploration** lets π_{explore} act directly in the 102-dimensional joint space without use of a learned prior. **Hand-Only Prior** is similar to Omnigrasp (Luo et al., 2024a) and PULSE (Luo et al., 2024b) which lets π_{explore} act in a latent space distilled from experts that track hand motion without the object.



(a) Success rate (b) Hand usage (c) Latent space (PCA)

Figure 4: **Effect of the latent residual bound** B. Exploration with less constraint (larger bound value B) leads to higher success rate (a) but collapsed grasp diversity (b) as the deployed latents (blue) deviate too much from the learned latent prior (grey) in (c).

Results. To accelerate baseline learning, we additionally give both baselines a pre-grasp reward (Dasari et al., 2023; Luo et al., 2024a) and a difficulty curriculum (Section D.3). Our method needs neither and is trained on the full task distribution from the start. As Table 2 shows, it nonetheless outperforms both baselines on unseen objects, improving the success rate over the strongest baseline by 14.0 points on GRAB and 16.9 points on DexYCB while also reducing object pose errors. It also reaches the peak baseline performance with only 1/10 of the training time (in 900 epochs instead of 8900 epochs), demonstrating much higher training efficiency. Because the DexYCB policy builds on the prior trained exclusively on GRAB, these results further show that the interaction prior generalizes across datasets. Qualitative results from both datasets are depicted in Figure 3.

Table 2: **Tabletop manipulation synthesis** on GRAB and DexYCB with unseen objects and trajectories. The interaction prior is trained only on GRAB dataset.

	GRAB (Taheri et al., 2020)				DexYCB (Chao et al., 2021)			
Method	SR ↑	TC ↑	E_t (cm)↓	E_r (°)↓	SR ↑	TC ↑	E_t (cm)↓	E_r (°)↓
Action Space Expl.	59.1	73.6	1.79	21.5	60.8	81.3	1.63	14.2
Hand-Only Prior	60.0	73.4	1.52	21.2	59.5	79.4	1.59	16.1
Ours	**74.0**	**82.9**	**1.36**	**20.0**	**77.7**	**87.4**	**1.52**	**13.8**

4.3 Generalization to New Tasks

The tabletop manipulation policy above is trained on similar interactions to those the prior was trained on. We now ask whether the interaction prior transfers to tasks whose dynamics are absent or very rare in the training data. We consider two: *catching* a thrown object, and a bimanual *handover* in which the object must be passed from the left hand to the right. Both are formulated as object-trajectory-conditioned manipulation generation, so the exploration policy π_{explore}, the object split, and the baselines are unchanged from Section 4.2. The object-tracking reward is complemented by simple task-specific terms. We show success rate in Tab. 3 and qualitative examples in Fig. 5. Task definitions, implementation details and full result tables are given in Appendices E.3–E.5.

Table 3: Success rate (%↑) on held-out GRAB objects for new tasks.

Method	Catching	Handover
Action Space Expl.	43.0	11.9
Hand-Only Prior	1.0	5.0
Ours (right hand)	78.0	–
Ours	**87.0**	**51.9**

Catching. An object with random orientation and spin is launched on a ballistic arc towards an intercept point. The hands start open, palms facing, at a random offset (Sec. E.4). We define both the bimanual- and right-hand catching task, where we encourage different hands to catch the thrown object. Tab. 3 shows that our interaction prior is highly beneficial, achieving 2.0× the success rate of the strongest baseline.

Handover. This task is designed to force a handover of the object from the left to the right hand. The object is initialized in the left hand, and the reference trajectory moves it in a straight line to the right side of the workspace. Each hand is allowed to stay only half of the workspace, ensuring that object movement is only possible by a proper change of hands. Tab. 3 shows that our interaction prior is critical for this challenging task with $4.3\times$ the success rate of the strongest baseline.

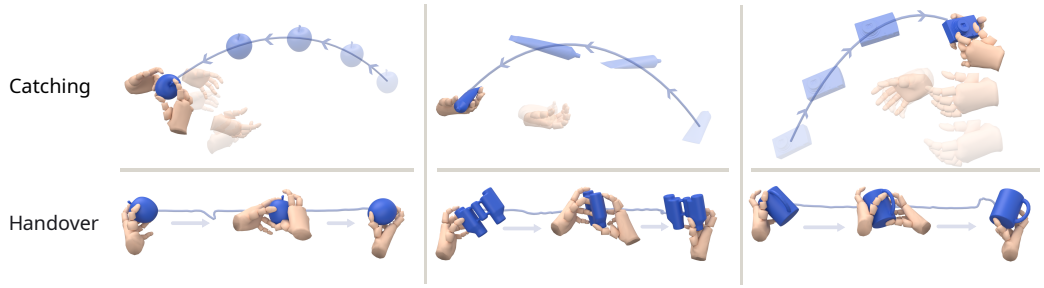


Figure 5: **Generalization of the interaction prior** to object catching (top) and handover (bottom). Our interaction prior transfers to other manipulation tasks rarely seen in GRAB.

4.4 Ablations

We run all ablations on Tabletop Manipulation on GRAB.

Stage III components. Table 4 ablates three components of Stage III, which we define formally and show full table in Appendix F.3. Executing the frozen prior without an exploration policy reaches only 30.8% success, so Stage III more than doubles the performance of the distilled prior. Adding the residual to the decoded action instead of the latent lowers the success to 49.6%, showing that exploring within the prior's latent space is substantially more effective than correcting its actions. Random trajectory augmentation also boosts the success rate by about six points.

Table 4: Stage III ablations, success rate on held-out GRAB objects.

Method	SR ↑
Direct prior rollout	30.8
Action-space residual	49.6
w/o random traj.	67.9
Ours	**74.0**

Residual bound. We introduce a bound in the residual latent to constrain the RL exploration in stage III. We analyze how this affects success rate and grasp diversity in Fig. 4. Success rises as the bound is relaxed and up to $B{=}5$ the deployed latents stay within the prior's distribution and hand usage follows the human reference. Looser bounds improve success only marginally, but the latents drift away from the prior and the policy collapses almost entirely to bimanual grasps, so we use $B{=}5$.

Robustness. Appendix F.1 evaluates robustness to changes in friction, object mass, and action noise. Our policy retains the highest success rate under all perturbations, and degrades far less than the baselines when physical object properties deviate.

5 Conclusion

We propose PiHand, a physics-based framework for synthesizing dexterous hand–object interactions from target object trajectories. Our key idea is to learn an explicit interaction prior that jointly models hand motion, object geometry, and contact dynamics, rather than relying on hand-only priors or reward-driven contact discovery. PiHand trains imitation experts on mocap data, distills them into a conditional VAE prior, and explores its latent space to solve novel tasks. Our experts improve SR over ManipTrans from 69.6% to 93.3% and maintain 84.8% SR on the full GRAB set with per-object experts. On unseen GRAB and DexYCB objects, our tabletop manipulation policies achieve 14.0 and 16.9 points higher SR than the strongest baselines while learning roughly ten times faster. The same prior transfers to catching and handover, tasks rare or absent in its training data, achieving $2.0\times$ and $4.3\times$ the strongest-baseline SR, respectively. We will release the processed data, expert policies, and generation model.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing parts of our method. We did not use generative AI tools to develop theoretical models, formulate or prove mathematical claims, propose hypotheses, design experiments, interpret results, or for synthetic data generation, translation, dataset cleaning, or qualitative data analysis. Additionally, we used generative AI tools to draft and rewrite parts of the paper, improve its readability, suggest a reorganization of the manuscript, help with table and general LaTeX-formatting, help with finding related work, and write Blender scripts for our visualizations. All AI-generated code was reviewed and tested by the authors, all AI-assisted text was checked and revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. Section 3 describes all three stages of PiHand. Appendix B complements it with the network architectures, the hyperparameters (Table 5), the observation horizons and the termination criteria of each stage. Both datasets we use, GRAB and DexYCB, are publicly available. Appendix C.1 explains how we preprocess them. Appendix C.2 describes the simulated MANO-based hand and all physics parameters, and Appendix C.3 reports hardware and training times. Baseline implementations and their additional rewards and curriculum are described in Appendix D. The Stage III task definitions are documented in Appendix E. We will release the full code, the processed data, the expert policies and the model checkpoints upon publication.

REFERENCES

Jianibieke Adalibieke, Qianwei Han, Xueyi Liu, Yuzhe Qin, and Li Yi. Adadextrack: Dynamic modulation for adaptive and generalizable dexterous manipulation tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28021–28031, June 2026.

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. Visual imitation enables contextual humanoid control. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 794–815. PMLR, 2025. URL https://proceedings.mlr.press/v305/allshire25a.html.

Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis, 2023. URL https://arxiv.org/abs/2309.07907.

Junuk Cha, Jihyeon Kim, Jae Shin Yoon, and Seungryul Baek. Text2hoi: Text-guided 3d motion generation for hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1577–1585, 2024.

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

Zerui Chen, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Vividex: Learning vision-based dexterous manipulation from human videos. *arXiv preprint arXiv:2404.15709*, 2024. URL https://arxiv.org/abs/2404.15709.

Zixuan Chen, Mazeyu Ji, Xuxin Cheng, Xuanbin Peng, Xue Bin Peng, and Xiaolong Wang. Gmt: General motion tracking for humanoid whole-body control, 2025. URL https://arxiv.org/abs/2506.14770.

Sudeep Dasari, Abhinav Gupta, and Vikash Kumar. Learning dexterous manipulation from exemplar object trajectories and pre-grasps. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3889–3896, 2023. doi: 10.1109/ICRA48891.2023.10161147.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8944–8951, 2024. doi: 10.1109/IROS58592.2024.10801984. URL https://human2humanoid.com/.

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, Linxi Fan, and Yuke Zhu. HOVER: Versatile neural whole-body controller for humanoid robots. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025. URL https://arxiv.org/abs/2410.21229. arXiv:2410.21229.

Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024. URL https://arxiv.org/abs/2412.13196.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: Automated data generation for bimanual dexterous manipulation via imitation learning. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025. URL https://dexmimicgen.github.io/.

Hwanhee Jung, Seunggwan Lee, Jeongyoon Yoon, SeungHyeon Kim, Giljoo Nam, Qixing Huang, and Sangpil Kim. Decoupled generative modeling for human-object interaction synthesis, 2025. URL https://arxiv.org/abs/2512.19049.

Dvij Kalaria, Sudarshan S Harithas, Pushkal Katara, Sangkyung Kwak, Sarthak Bhagat, Shankar Sastry, Srinath Sridhar, Sai Vemprala, Ashish Kapoor, and Jonathan Chung-Kuan Huang. Dreamcontrol: Human-inspired whole-body humanoid control for scene interaction via guided diffusion, 2025. URL https://arxiv.org/abs/2509.14353.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis. *ACM Transactions on Graphics*, 42(6), 2023. doi: 10.1145/3618333. URL https://lijiaman.github.io/projects/omomo/.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C. Karen Liu. Controllable human-object interaction synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. doi: 10.48550/arXiv.2312.03913. URL https://lijiaman.github.io/projects/chois/.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. Maniptrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025a.

Muchen Li, Sammy Christen, Chengde Wan, Yujun Cai, Renjie Liao, Leonid Sigal, and Shugao Ma. Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17416–17425, June 2025b.

Sikai Li, Shuning Li, Zhenyu Wei, Yunchao Yao, Chenran Li, and Mingyu Ding. Coordex: Coordinating body and hand priors for continuous dexterous humanoid loco-manipulation, 2026. URL https://arxiv.org/abs/2606.23680.

Xiaotong Lin, Tianming Liang, Jian-Fang Hu, Kun-Yu Lin, Yulei Kang, Chunwei Tian, Jianhuang Lai, and Wei-Shi Zheng. Coopdiff: Anticipating 3d human-object interactions via contact-consistent decoupled diffusion, 2025. URL https://arxiv.org/abs/2508.07162.

Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. Dextrack: Towards generalizable neural tracking control for dexterous manipulation from human references, 2025a. URL https://arxiv.org/abs/2502.09614.

Xueyi Liu, He Wang, and Li Yi. Dexndm: Closing the reality gap for dexterous in-hand rotation via joint-wise neural dynamics model, 2025b. URL https://arxiv.org/abs/2510.08556.

Zhengyi Luo, Jinkun Cao, Alexander W. Winkler, Kris Kitani, and Weipeng Xu. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/papers/Luo_Perpetual_Humanoid_Control_for_Real-time_Simulated_Avatars_ICCV_2023_paper.pdf.

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 2161–2184, 2024a. doi: 10.52202/079017-0070.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://arxiv.org/abs/2310.04582. arXiv:2310.04582.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance gpu-based physics simulation for robot learning, 2021. URL https://arxiv.org/abs/2108.10470.

Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. Dexmachina: Functional retargeting for bimanual dexterous manipulation, 2025. URL https://arxiv.org/abs/2505.24853.

Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.19901.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985, 2019.

Xiaogang Peng, Yiming Xie, Zizhao Wu, Varun Jampani, Deqing Sun, and Huaizu Jiang. Hoi-diff: Text-driven synthesis of 3d human-object interactions using diffusion models. *arXiv preprint arXiv:2312.06553*, 2023.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4):1–14, 2018.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics*, 41(4), 2022. URL https://research.nvidia.com/labs/toronto-ai/ASE/.

Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. Dexmv: Imitation learning for dexterous manipulation from human videos, 2021. URL https://arxiv.org/abs/2108.05877.

Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations, 2017. URL https://arxiv.org/abs/1709.10087. Accepted to RSS 2018.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6): 245:1–245:17, November 2017. doi: 10.1145/3130800.3130883. URL http://doi.acm.org/10.1145/3130800.3130883.

Stéphane Ross, Geoffrey J. Gordon, and J. Andrew 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 (AISTATS)*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv e-prints*, pp. arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. URL https://grab.is.tue.mpg.de.

Omid Taheri, Vasileios Choutas, Michael J. Black, and Dimitrios Tzionas. GOAL: Generating 4D whole-body motion for hand-object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13253–13263, June 2022. doi: 10.1109/CVPR52688.2022.01291. URL https://goal.is.tue.mpg.de/.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 2024. doi: 10.1145/3687951. URL https://arxiv.org/abs/2409.14393.

Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. Masked-Manipulator: Versatile whole-body control for loco-manipulation. In *SIGGRAPH Asia Conference Papers*, 2025. doi: 10.1145/3757377.3763934. URL https://arxiv.org/abs/2505.19086. arXiv:2505.19086.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-based imitation of dynamic human-object interaction. arXiv preprint, 2023. URL https://arxiv.org/abs/2312.04393.

Yinhuai Wang, Runyi Yu, Hok Wai Tsui, Xiaoyi Lin, Hui Zhang, Qihan Zhao, Ke Fan, Miao Li, Jie Song, Jingbo Wang, Qifeng Chen, and Ping Tan. Learning generalizable hand-object tracking from synthetic demonstrations. arXiv preprint, 2025a. URL https://arxiv.org/abs/2512.19583.

Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, Jian Zhang, Lei Zhang, and Ping Tan. SkillMimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Wang_SkillMimic_Learning_Basketball_Interaction_Skills_from_Demonstrations_CVPR_2025_paper.pdf.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3D meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics (TOG)*, 41 (4):1–18, 2022. doi: 10.1145/3528223.3530103.

Zhen Wu, Jiaman Li, Pei Xu, and C. Karen Liu. Human-object interaction from human-level instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11176–11186, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wu_Human-Object_Interaction_from_Human-Level_Instructions_ICCV_2025_paper.html.

Weiji Xie, Jinrui Han, Jiakun Zheng, Huanyu Li, Xinzhe Liu, Jiyuan Shi, Weinan Zhang, Chenjia Bai, and Xuelong Li. Kungfubot: Physics-based humanoid whole-body control for learning highly-dynamic skills, 2025. URL https://arxiv.org/abs/2506.12851.

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. InterMimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2502.20390. CVPR 2025 Highlight; arXiv:2502.20390.

Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. Interprior: Scaling generative control for physics-based human-object interactions, 2026. URL https://arxiv.org/abs/2602.06035.

Shaofeng Yin, Yanjie Ze, Hong-Xing Yu, C. Karen Liu, and Jiajun Wu. Visualmimic: Visual humanoid loco-manipulation via motion tracking and generation, 2025. URL https://arxiv.org/abs/2509.20322.

Dongjie Yu, Kun Lei, Zhennan Jiang, Jia Pan, and Huazhe Xu. Beyond action residuals: Real-world robot policy steering via bottleneck latent reinforcement learning, 2026. URL https://arxiv.org/abs/2605.19919.

Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. *arXiv preprint arXiv:2505.02094*, 2025. arXiv:2505.02094.

Yanjie Ze, Zixuan Chen, João Pedro Araújo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C. Karen Liu. Twist: Teleoperated whole-body imitation system. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 2143–2154. PMLR, 2025. URL https://proceedings.mlr.press/v305/ze25a.html.

Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592408.

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. Graspxl: Generating grasping motions for diverse objects at scale. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, volume 15084 of *Lecture Notes in Computer Science*, pp. 386–403. Springer, Cham, 2025a. doi: 10.1007/978-3-031-73347-5_22.

Wanyue Zhang, Rishabh Dabral, Vladislav Golyanik, Vasileios Choutas, Eduardo Alvarado, Thabo Beeler, Marc Habermann, and Christian Theobalt. Bimart: A unified approach for the synthesis of 3d bimanual interaction with articulated objects. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Vladimir Guzov, and Gerard Pons-Moll. Couch: Towards controllable human-chair interactions. In *European Conference on Computer Vision (ECCV)*. Springer, October 2022. URL https://virtualhumans.mpi-inf.mpg.de/couch/.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez Pellitero, and Gerard Pons-Moll. Force: Dataset and method for intuitive physics guided human-object interaction. *arXiv preprint arXiv:2403.11237*, 2024. URL https://virtualhumans.mpi-inf.mpg.de/force/.

Zhikai Zhang, Haofei Lu, Yunrui Lian, Ziqing Chen, Yun Liu, Chenghuai Lin, Han Xue, Zicheng Zeng, Zekun Qi, Shaolin Zheng, Qing Luan, Jingbo Wang, Junliang Xing, He Wang, and Li Yi. Learning athletic humanoid tennis skills from imperfect human motion data, 2026. URL https://arxiv.org/abs/2603.12686.

Siheng Zhao, Yanjie Ze, Yue Wang, C. Karen Liu, Pieter Abbeel, Guanya Shi, and Rocky Duan. Resmimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning, 2025. URL https://arxiv.org/abs/2510.05070.

Chunxin Zheng, Kai Chen, Zhihai Bi, Yulin Li, Liang Pan, Jinni Zhou, Haoang Li, and Jun Ma. Embracing bulky objects with humanoid robots: Whole-body manipulation with reinforcement learning, 2025. URL https://arxiv.org/abs/2509.13534.

Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Gears: Local geometry-aware hand-object interaction synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20634–20643, June 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhou_GEARS_Local_Geometry-aware_Hand-object_Interaction_Synthesis_CVPR_2024_paper.html.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5745–5753, 2019. doi: 10.1109/CVPR.2019.00589. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhou_On_the_Continuity_of_Rotation_Representations_in_Neural_Networks_CVPR_2019_paper.html.

A LIMITATIONS

While PiHand clearly improves over prior baselines, several limitations remain. First, our tracking experts do not yet reach 100% success rate on the full GRAB dataset. Failures concentrate on extremely dynamic motions and sequences with heavy mocap noise, which we expect could be addressed with stricter termination criteria, curriculum learning, or higher-quality reference data. Second, we focus on rigid objects and do not model articulated objects such as those in ARCTIC (Fan et al., 2023). Extending the interaction prior to articulated geometry is an important next step. Third, our framework controls only the hands and does not consider full-body motion; combining a full-body locomotion controller with our hand manipulation prior is a promising future direction. Finally, all experiments are conducted in simulation with a single MANO-based hand model, whose self-collisions within each hand are only approximated by a penetration check (Appendix B.1). Transferring PiHand to real robotic hands would be an exciting direction, but it requires retargeting the prior to a different embodiment and bridging the gap between simulated and real sensing and actuation, which we consider a substantial project in its own right.

B IMPLEMENTATION DETAILS

This section provides the network architectures, hyperparameters and termination criteria needed to reproduce the three stages of our pipeline, starting with details shared by all stages.

B.1 SHARED DETAILS

Observation horizon. In all stages, the observation contains future frames of the object reference trajectory and, where applicable, future mocap hand poses. Stages I and III sample these frames at the offsets $\mathcal{T} = \{1, 2, 3, 4, 6, 8, 12, 16, 24, 32, 48, 64, 96, 128\}$. To avoid overfitting of the latent prior, Stage II uses the reduced set $\mathcal{T} = \{1, 2, 4, 8, 16, 32, 64\}$.

Self-collisions. In IsaacGym we disable all collisions between the rigid bodies within one hand model. The simulator cannot identify rigid-body contact pairs and only provides the total contact force acting on each rigid body. Since we infer the binary contact maps $\mathbf{c}_t$ from these contact forces, we must ensure that they stem from the hand colliding with the scene and not with itself. To prevent the resulting unfiltered self-penetrations of the hand, we check for self-penetration by approximating the fingers as cylinders of 1 cm diameter, and use these checks as a termination criterion in all stages of the pipeline.

B.2 STAGE I: EXPERT TRAINING

Keypoint weighting. The keypoint loss L^{key} tracks all $2n_K$ keypoints in a frame that interpolates between wrist-local (non-contact) and object-centric (in-contact) via the time-to-contact weight $w_i = \exp(-\tau_i/\tau_0)$, where τ_i is the number of frames until keypoint i next contacts the object in the reference mocap trajectory, and $\tau_0 = 30$.

Hard-mining sampling. We sample reference sequence i from the expert's pool with a probability that interpolates between uniform and failure-proportional sampling:

$$P(i) = (1 - \alpha)\tfrac{1}{N} + \alpha\, \tfrac{1-\text{TC}_i+\epsilon}{\sum_{j=1}^{N}(1-\text{TC}_j+\epsilon)}, \tag{1}$$

with pool size N, recent tracking-completion rate $\text{TC}_i \in [0, 1]$, hard-mining strength α, and numerical-stability constant $\epsilon = 10^{-6}$. This ensures that harder sequences are sampled more frequently than those that are already tracked well.

Networks and optimization. The actor is a 3-layer MLP with hidden layer sizes (1024, 1024, 512), and the critic is a 2-layer MLP with layer sizes (1024, 512). Both networks use GELU activations. All PPO hyperparameters are summarized in Table 5.

B.3 STAGE II: DISTILLATION

Encoder, prior and decoder are MLPs with layer sizes (2048, 2048, 1024, 1024, 512), GELU activations and layer normalization, and share a latent space of dimension 64. All spatial observations,

Table 5: Hyperparameters for expert policy training. Reward weights and environment thresholds are shared across all expert variants. We scale the minibatch size with the dataset diversity of the respective training task.

Parameter	**Value**
Reward Weights & Environment	
Keypoint tracking weight (λ_{key})	10
Object surface-sample weight (λ_{samp})	10
Contact tracking weight (λ_{cont})	2.5
Termination threshold: Object position deviation	5 cm $\rightarrow$ 2 cm (annealed)
Termination threshold: Object sample deviation	5 cm $\rightarrow$ 2 cm (annealed)
Action standard deviation	Learnable, initialized at $\exp(-1)$
Optimization & PPO Settings	
Optimizer	Adam
Learning rate	2×10^{-5}
Number of environments	4096
Rollout length (steps per env)	32
Discount factor (γ)	0.99
GAE parameter (τ)	0.95
Clip range (e_{clip})	0.2
PPO mini-epochs	6
Minibatch size (ManipTrans comparison)	1024
Minibatch size (Per-object experts)	16384

such as the object pose, are passed to these networks expressed in the wrist frames of *both* hands, which makes them invariant to global scene transformations. Following Luo et al. (2024a), we parameterize the encoder mean as a residual on the prior mean, $\mu_{\mathcal{E}} = \mu_{\mathcal{P}}(\tilde{\mathbf{o}}_t) + \Delta\mu(\mathbf{o}_t)$, where the encoder network outputs the residual $\Delta\mu(\mathbf{o}_t)$. We linearly anneal the KL loss weight from 0 to 10^{-4} and additionally apply episodic latent noise (Tessler et al., 2024) to prevent posterior collapse. All networks are trained with Adam and a learning rate of 10^{-4}. To avoid overfitting of the prior to the training sequences, we remove all velocities from the simulation state observation $\mathbf{o}_t^{\text{sim}}$, represent future reference poses only relative to the current state, and use a shorter look-ahead (see Appendix B.1). This improved the direct prior rollout success rate on the held-out evaluation set from 23.7% to 30.8%, when compared with a prior trained on full observations.

The simulation setup is identical to Stage I, and we employ the same hard-mining strategy. We terminate environments early when the object position deviates by more than 5 cm from the reference, since the experts never encountered larger deviations during their own training and their supervision would be unreliable in such states.

B.4 Stage III: Latent Exploration

The latent exploration policy is an MLP with hidden layer sizes (2048, 1024, 1024, 512) and ReLU activations, and we bound its latent residual with $B = 5$. For Tabletop Manipulation, the only reward term is $\exp(-10\, L^{\text{samp}})$, based on the object surface-sample loss from Stage I; the rewards for the downstream tasks are described in Section E.3. Episodes terminate early when the object deviates by more than 10 cm from its reference position, which matches the evaluation threshold. The action standard deviation is initialized to 1, and all remaining optimization and PPO settings are the same as in Stage I.

C Data and Simulation

C.1 Datasets and Preprocessing

GRAB. We use the GRAB (Taheri et al., 2020) dataset as the primary source of interaction demonstrations. GRAB contains high-quality whole-body human grasping motions with detailed hand poses and contact annotations, collected from 10 subjects interacting with 51 everyday objects. We use 50 objects in our experiments, excluding `doorknob` due to physically implausible floating trajectories in the source data. Following prior work (Taheri et al., 2022), we partition the dataset into a training and a cross-object evaluation split. Raw motion capture sequences often contain extensive segments without actual object manipulation, such as the initial arm approach or the post-release retraction. To focus our policies on the core interaction dynamics, we temporally trim the sequences to segments that begin approximately 40 frames before the initial contact and end 40 frames after the contact is broken. Finally, to prevent lateral bias and increase data diversity, we apply a left-right mirroring augmentation to all sequences. After preprocessing, the final dataset contains 4,854 training sequences with an average length of 317.4 frames at 60 Hz.

DexYCB. We use DexYCB (Chao et al., 2021) to evaluate generalization beyond GRAB. DexYCB contains sequences of 10 subjects grasping 20 objects from the YCB object set with a single hand, annotated with MANO hand poses and 6D object poses. In each sequence, a subject picks up a single target object from a table among several distractor objects and lifts it. We use the 250 right-handed sequences of subjects 1–5, covering all 20 objects. Following the unseen-object setting (S3) of DexYCB, we train on only 15 objects (186 sequences) and hold out 5 objects for evaluation (`005_tomato_soup_can`, `009_gelatin_box`, `021_bleach_cleanser`, `036_wood_block`, `061_foam_brick`; 64 sequences). As for GRAB, we temporally trim the sequences to the interaction. Because all sequences are right-handed, we embed them into our bimanual representation with the left hand idle and do not apply mirroring augmentation. After preprocessing, the sequences have an average length of 196.4 frames at 60 Hz.

C.2 Simulation Setup and MANO Integration

To bridge MANO-based human demonstrations and physical simulation, we construct our simulated robotic hand directly from the MANO model (Romero et al., 2017). The simulated hand follows the kinematic joint structure of the mean-shape MANO template in a flat-hand pose, with rigid-body collision geometries obtained by segmenting the corresponding MANO mesh. This design preserves a direct correspondence between the simulated and the visual hand representation, so that simulated trajectories can be mapped back to MANO without error-prone kinematic retargeting. While the GRAB dataset contains motions from 10 different human subjects with varying hand shapes, we use this single template hand across all environments. Rather than retargeting each demonstration to the template, we let the Stage I tracking policy learn to compensate for the remaining morphology mismatch during RL training.

For all simulated objects, we set the volumetric density to 500 kg/m^3 and the contact friction coefficient to 0.7. Prior methods often use elevated friction values that artificially simplify grasping, such as 1.0 in Omnigrasp (Luo et al., 2024a) and 1.5 in MaskedManipulator (Tessler et al., 2025). Our lower friction prevents unrealistic surface adhesion and encourages the policy to learn physically plausible force-closure grasps. Simulation and policy execution are synchronized at 60 Hz. Since IsaacGym assumes convex meshes for efficient collision detection, we decompose all object meshes into convex parts using Wei et al. (2022).

C.3 Training Times

All models are trained on a single Nvidia A100. Each Stage I expert is trained for up to 6 h, however, on average they converge after only 1 or 2 h. Distillation into the latent interaction prior takes 24 h. Training the Tabletop Manipulation policy takes up to 36 h, while the catching and handover policies are trained for 20 h.

D BASELINES

Several existing methods qualify as baselines for generative hand–object interaction, most notably full action space exploration frameworks such as GraspXL (Zhang et al., 2025a), and human-motion latent exploration methods such as Omnigrasp (Luo et al., 2024a) and the work Braun et al. (2023).

However, conducting fair, direct comparisons using official codebases is challenging for several reasons. First, some methods lack publicly available or well-documented code, making reproduction outside of their original environment practically impossible without major engineering efforts. Second, existing methods leverage vastly different training datasets. Crucially, differences in the underlying physics simulation—such as the choice of engine, friction coefficients, object mass densities, and the specific embodiment—fundamentally alter the difficulty of the task. For instance, full-body methods (Braun et al., 2023; Luo et al., 2024a) approximate an SMPL-X (Pavlakos et al., 2019) body using coarse geometric primitives (capsules and cuboids), leading to radically different contact dynamics than high-fidelity, MANO-based hand interactions.

To isolate the impact of the interaction-first prior and ensure a strictly fair comparison, we implement the paradigms of these baselines directly within our framework. By standardizing the simulator, physics parameters, and embodiment across all methods, any observed difference in performance can be attributed to the approaches themselves rather than to environmental discrepancies. In the following we describe the concrete details of our baseline implementations.

D.1 ACTION SPACE EXPLORATION

The action space exploration baseline represents the unconstrained RL approach. Here, the policy π_{explore} directly outputs continuous actions in the 102-dimensional PD-target residual space. No generative prior is learned from human demonstrations; instead, the policy must discover grasping mechanics relying purely on the task-specific reward signal. This mirrors the exploration paradigm of frameworks such as GraspXL (Zhang et al., 2025a).

D.2 HAND-ONLY PRIOR

Methods like Omnigrasp (Luo et al., 2024a) improve upon raw action space exploration by first compressing human imitation skills into a proprioceptive, motion-only latent space, which is subsequently explored to solve object interactions. We reimplement this paradigm as our Hand-Only Prior baseline. To distinguish the observations in this motion-only setting from our interaction-aware setting $(\mathbf{o}_t, \tilde{\mathbf{o}}_t)$, we denote them as $\mathbf{x}_t$.

D.2.1 HAND-ONLY IMITATION EXPERTS

We first train tracking experts exclusively on the hand kinematics of the GRAB dataset, entirely ignoring the object. The main differences to our interaction expert training are outlined below.

Observation space. The policy observation is $\mathbf{x}_t = (\mathbf{x}_t^{\text{sim}}, \mathbf{x}_t^{\text{ref}}, \mathbf{a}_{t-1})$. Because the environment contains neither a table nor a target object, the state dimensions are significantly reduced. The simulation state consists only of the hand root poses, joint angles, and keypoints and their velocities, $\mathbf{x}_t^{\text{sim}} = (\mathbf{q}, \dot{\mathbf{q}}, \mathbf{K}, \dot{\mathbf{K}})$, and the reference observation omits the object trajectory, $\mathbf{x}_t^{\text{ref}} = \left\{\hat{\mathbf{q}}_\tau,\ \Delta\hat{\mathbf{q}}_\tau,\ \hat{\mathbf{K}}_\tau,\ \hat{\mathbf{K}}_\tau - \mathbf{K}\right\}_{\tau\in\mathcal{T}}$.

Reward formulation. The policy is rewarded solely for minimizing the distance to the reference keypoints, using the single reward term $\exp\big(-\frac{\lambda}{2n_K}\sum_{j=1}^{2n_K}\|\mathbf{K}_j - \hat{\mathbf{K}}_j\|_2\big)$ with $\lambda = 100$.

Training details. Without contact dynamics to model, this imitation task is much simpler, so we do not require per-object experts. Instead, we partition the dataset into 4 subsets and train one expert per subset. Episodes terminate early upon hand self-penetration or if the average keypoint error exceeds 2.5 cm, and we use the same hard-mining strategy as for our interaction experts. As Table 6 shows, the hand-only experts achieve high success rates and a very low MPJPE even under this strict threshold, so the baseline prior is distilled from well-trained experts.

Table 6: Hand-only imitation experts used to build the Hand-Only Prior baseline. SR and TC in % under a 2.5 cm keypoint error threshold, MPJPE in cm.

	Dataset	SR ↑	TC ↑	MPJPE ↓
Hand-only imitation experts	GRAB + mirroring	79.5	84.5	0.53

D.2.2 Distillation into a Hand-Only Latent Space

We keep the conditional VAE architecture with a learnable prior, but remove all object-related information from its inputs. The partial observation is simply the simulated hand state, $\tilde{\mathbf{x}}_t = \mathbf{x}_t^{\text{sim}}$, while the privileged observation passed to the encoder during training additionally contains the future reference hand kinematics, $\mathbf{x}_t = (\tilde{\mathbf{x}}_t, \mathbf{x}_t^{\text{ref}})$. Both the prior and the decoder are conditioned only on $\tilde{\mathbf{x}}_t$. The distillation loss and KL regularization are identical to our interaction prior.

D.2.3 Latent Exploration

For the generative tasks, we add the table and the object to the scene. The exploration policy π_{explore} therefore needs to know the object and its target trajectory, so it receives the object-aware partial observation $\tilde{\mathbf{o}}_t$ of our main method rather than the restricted $\tilde{\mathbf{x}}_t$, and predicts a residual $\Delta\mathbf{z}_t = \pi_{\text{explore}}(\tilde{\mathbf{o}}_t)$. Because the pre-trained decoder remains blind to the object, this residual is added to the motion-only prior mean and decoded from the proprioceptive state alone, $\mathbf{a}_t = \mathcal{D}\big(\tilde{\mathbf{x}}_t, \mu_{\mathcal{P}}(\tilde{\mathbf{x}}_t) + \Delta\mathbf{z}_t\big)$. This restricts exploration to human-like motions but deprives the decoder of any explicit interaction awareness.

D.3 Additional Rewards and Curriculum for Tabletop Manipulation

For Tabletop Manipulation, our method is trained with the object surface-sample reward alone and directly on the full task distribution. The baselines have no prior exposure to objects, and with this setup the initially uninformative tracking signal rarely leads the hands to reach the object or establish a stable grasp; in our experiments, neither baseline learned meaningful behavior within 6 h of training. To accelerate their learning, we add a pre-grasp reward and a training curriculum, as summarized in Table 7. Apart from these two additions, the baselines follow the training procedure of Section 3.3.

Table 7: Reward terms and curriculum used by each method for Tabletop Manipulation. Although our method relies only on the object surface-sample reward and needs no curriculum, it outperforms both baselines.

Method	Obj. Sample	Pre-grasp	Curriculum
Action Space Exploration	✓	✓	✓
Hand-Only Prior	✓	✓	✓
Ours	✓	×	×

Pre-grasp reward. Following Dasari et al. (2023); Luo et al. (2024a), we add a pre-grasp reward term. Since each episode initializes from a known training-pool state, we have access to the corresponding ground-truth pre-grasp hand pose $\hat{\mathbf{q}}^{\text{pg}}$ at the first contact frame. We use it *only* as a reward signal and never as a policy input, which preserves the zero-shot setting at inference. The pre-grasp loss L^{pregrasp} is the Euclidean keypoint distance between the simulated hands and $\hat{\mathbf{q}}^{\text{pg}}$. We split each episode into two sequential phases. During the initial approach phase, the reward only encourages reaching the pre-grasp pose,

$$r_t = \exp\big(-\lambda_{\text{grasp}}\, L^{\text{pregrasp}}\big),$$

and after a fixed number of initial steps (40 frames), it switches to the object tracking objective,

$$r_t = \exp\big(-\lambda_{\text{samp}}\, L^{\text{samp}}\big).$$

This explicit transition prevents the two objectives from interfering, so the policy first focuses on establishing a stable force closure before shifting its capacity to realizing the target trajectory.

Table 8: Three-seed replication of the ManipTrans comparison on GRAB50. We report mean ± standard deviation over 3 random seeds.

Method	Data	Granularity	SR (5cm) [%]	TC (5cm) [%]	E_t [cm]	E_r [°]	MPJPE [cm]
ManipTrans (Li et al., 2025a)	GRAB50	Per-sequence	69.6 ± 1.5	77.6 ± 1.9	0.43 ± 0.01	7.3 ± 0.2	1.97 ± 0.06
Ours	GRAB50	Per-sequence	$\mathbf{93.3 \pm 3.0}$	$\mathbf{95.8 \pm 2.6}$	$\mathbf{0.28 \pm 0.05}$	$\mathbf{4.9 \pm 1.2}$	$\mathbf{1.01 \pm 0.10}$

Curriculum. Direct training over the highly diverse initial poses and trajectories proves difficult for the baselines, so we introduce a three-level curriculum that progressively increases task complexity. In level 1, environments are initialized with matched hand and object poses drawn from the same training sequence. Level 2 adds continuous z-axis rotation augmentations to both the initial poses and the target trajectory. Level 3 completely decouples the initial configurations of hands and object by sampling the hand pose from a different training sequence than the object pose. Our method is trained on this highest level from the start.

E Experiment Details and Additional Results

E.1 Comparison with ManipTrans

To evaluate our expert training against ManipTrans (Li et al., 2025a), we sample a subset of 50 right-handed sequences from the GRAB dataset. Because ManipTrans relies on a distinct, predefined MANO-based robotic hand model, we first retarget the source mocap data to their embodiment to ensure that their tracking controllers function correctly. Our formulation, in contrast, absorbs these morphological discrepancies and does not need explicit kinematic retargeting.

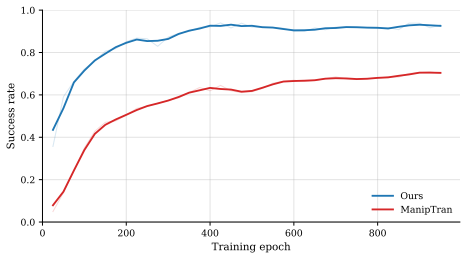


Figure 6: Success rate over training epochs of our imitation experts compared to ManipTrans on GRAB50.

Both methods are trained with identical hyperparameters for 1,000 epochs to ensure convergence, and evaluated with a strict 5 cm object translation threshold. The only difference to the per-object hyperparameters in Table 5 is the reduced minibatch size of 1024. Table 8 reports the results over three seeds. Figure 6 further shows that our method is considerably more sample efficient: it reaches the peak success rate of ManipTrans after only 175 epochs and ultimately converges to a clearly higher success rate and spatial precision.

E.2 Tabletop Manipulation

Table 9 reports the Tabletop Manipulation results on GRAB as mean and standard deviation over three random seeds. The gain over the baselines is consistent across seeds, and our method also shows by far the lowest variance.

Table 9: Tabletop Manipulation on held-out GRAB objects. We report mean ± standard deviation over 3 random seeds. SR and TC in % (10 cm threshold), E_t in cm, E_r in degrees.

Method	SR ↑	TC ↑	E_t ↓	E_r ↓
Action Space Exploration	59.1 ± 4.7	73.6 ± 3.6	1.79 ± 0.10	21.5 ± 1.0
Hand-Only Prior	60.0 ± 3.1	73.4 ± 2.1	1.52 ± 0.06	21.2 ± 1.0
Ours	$\mathbf{74.0 \pm 0.9}$	$\mathbf{82.9 \pm 0.6}$	$\mathbf{1.36 \pm 0.01}$	$\mathbf{20.0 \pm 0.5}$

E.3 Novel Tasks: Common Setup

Both novel tasks of catching and bimanual handover reuse the generative machinery of Sec. 4: a task is specified entirely by an object reference trajectory, and π_{explore} is optimized with an object-tracking reward, complemented by a small number of task-specific terms described below. This section defines the shared setup, and the following two sections define the reference trajectories, task-specific terminations, and the training and evaluation protocol of each task.

Scene. Neither task involves a physical table. The bimanual hands float in free space.

Reward. The reward is the exponentiated combination of an object position term (weight 10) and a fingertip-proximity term (weight 10 for catching, 5 for handover).

Termination. An episode terminates when the object leaves a task-specific radius around the reference (30 cm for catching, 15 cm for handover), when the nearest wrist is farther than a task-specific distance from the object (2.0 m and 1.2 m), or on self-penetration.

Objects. Catching uses all 45 training objects and the 5 held-out objects of Sec. 4. Handover requires a left-hand grasp frame to initialize from; five training objects (`knife`, `scissors`, `teapot`, `toothbrush`, `torussmall`) have none in GRAB and are excluded, leaving 40. All 5 held-out objects have left-hand grasp frames in the evaluation split.

E.4 Catching

Reference trajectory. The object is launched on a ballistic arc that passes through an intercept point, and the reference then holds the object at that point for at least 60 frames. To make the reference physically exact, the arc is integrated with the simulator's own symplectic Euler scheme rather than the closed-form parabola. After k substeps the position is

$$\mathbf{p}_k = \mathbf{p}_0 + k\,\Delta t\,\mathbf{v}_0 + \mathbf{g}\,\Delta t^2\,k(k{+}1)/2,$$

and the launch velocity is solved from this expression. Objects are given a random initial orientation and a random angular velocity of up to 8 rad/s; because the simulator rotates rigid bodies about the center of mass, the reference is computed for the center of mass and mapped back to the object origin using the per-object inertial offset. Episodes last 160 frames (2.67 s at 60 Hz).

Randomization. The intercept point is displaced from the nominal catch point by up to $(\pm 0.35, \pm 0.25, \pm 0.30)$ m, the hands start at a common random offset of up to $(\pm 0.15, \pm 0.10, \pm 0.10)$ m, the launch position lies 0.3–1.0 m from the intercept point at an azimuth of $\pm 75^\circ$ and a height of -0.3 to $+0.6$ m relative to it, and flight time is 15–70 frames (0.25–1.17 s). Both hands start open, palms facing each other 12 cm either side of the nominal catch point.

Wrist economy. With the tracking reward alone, interaction prior-based policies learn to send one hand out to the object early and ride along with it for the entire flight, a behavior the prior's data encourages (a GRAB hand is always near the object) and that the fingertip-proximity term pays for on every flight frame. We add two penalties on the wrists' linear and angular speed (weights 2.0 and 0.3), applied only before the object reaches the intercept point, and scale the fingertip-proximity term to 10% of its weight during the same window.

Right-hand catching. In this variant only the right hand may touch the object. Contact between the left hand and the object terminates the episode, and the fingertip-proximity and wrist-economy terms are computed for the right hand only. Because a single hand covers a smaller workspace, throws are narrower and flatter. The launch position lies 0.3–0.8 m from the intercept point at an azimuth of $\pm 50^\circ$ and a height of -0.2 to $+0.45$ m relative to it, and the intercept point is displaced from the nominal catch point by up to $(\pm 0.25, \pm 0.18, \pm 0.22)$ m. All other settings are identical to bimanual catching.

Training and evaluation. All policies are trained for 20h on a single A100. All methods receive the same wall-clock budget; because action space exploration does not evaluate a decoder at every step, it completes more updates in this time. Policies are evaluated deterministically with 20 episodes per object, giving 900 episodes on the 45 training objects and 100 on the 5 held-out objects. Throws are generated from a fixed seed, so every method is evaluated on the same sequence of objects, launch states and intercept points.

Results. Table 10 reports the full comparison. On the held-out objects, our bimanual policy catches the object in 87.0% of the episodes, compared with 43.0% for action space exploration, while the hand-only prior almost never succeeds. The picture is the same on the training objects, and our policy performs as well on the held-out objects as on the training objects. Restricting the catch to the right hand lowers the success rate to 78.0%, but even this harder variant clearly outperforms both bimanual baselines.

Table 10: Catching a thrown object. *Left:* the 45 training objects, 900 episodes per method. *Right:* five held-out objects, 100 episodes per method. All methods are trained under an equal wall-clock budget. The *right only* row is a different task, catching with the right hand only (Sec. E.4), and is not directly comparable to the other rows.

		Training Objects			Held-Out Objects		
Method	Hand Usage	SR ↑	TC ↑	E_t ↓	SR ↑	TC ↑	E_t ↓
Action Space Exploration	both hands	51.1	68.2	3.19	43.0	61.9	3.27
Hand-Only Prior	both hands	4.9	37.3	3.70	1.0	32.3	3.55
Ours	right only	78.2	89.8	2.03	78.0	88.3	1.98
Ours	both hands	**84.6**	**91.6**	**1.66**	**87.0**	**94.5**	**1.38**

E.5 HANDOVER

Initial state. Each episode starts from a GRAB frame in which the left hand alone holds the lifted object: the object center is more than 10 cm above the table, at least four left-hand contact bodies and at most one right-hand contact body touch it, and its speed is below 0.2 m/s. Where the object has same-actor handover sequences (GRAB intent *offhand*), 70% of resets are drawn from the pre-transfer part of those sequences. The hands and object are translated so that the object starts on the left side of the workspace, $x \in [0.30, 0.45]$ m in the table frame, with the right hand at its GRAB pose relative to the object. Five training objects (`knife`, `scissors`, `teapot`, `toothbrush`, `torussmall`) have no such frame and are excluded, leaving 22,641 start frames over 40 training objects. The five held-out objects have 1,442 start frames in the evaluation split.

A GRAB frame copied into the simulator is kinematically correct but not necessarily force-closed: the object may slip out of the left hand before any policy acts. To separate the handover from the problem of establishing a stable grasp in the first place, we test every held-out start frame without a policy. The hands are held at their initial joint configuration and wrist pose by the PD controller, and we record the displacement of the object from its initial position. A frame is kept if the object moves less than 5 cm within the first 30 frames (0.5 s). Because the test involves no policy, it selects the same states for every method. Of the 1,442 held-out frames, 160 pass.

Reference trajectory. The reference holds the object in place for 30 frames (*settle*), during which terminations are masked so that the policy can establish its own grip. It then moves the object in a straight line at constant height to $x \in [-0.45, -0.30]$ m at 0.2–0.4 m/s, with 0.5 s velocity ramps and a 0.5 s dwell at the midpoint, and finally holds it for at least 60 frames. Episodes last 420 frames (7 s).

Wrist zones. The left wrist may not move right of $x = -0.08$ m and the right wrist may not move left of $x = +0.08$ m. The zones are enforced by clamping the wrist position targets, and an episode terminates if a wrist is pushed more than 2 cm past its zone. Since both zones exclude the far side of the workspace, no single hand can carry the object from start to goal: an episode that runs to completion has brought the object into the right hand's zone.

Per-hand proximity. The fingertip-proximity term takes the nearest fingertip of each hand separately and sums the two distances, so that the free hand receives a gradient towards the object instead of being masked by the hand already holding it.

Metrics. SR and TC are defined as before. The translation error E_t is the mean distance between object and reference from the end of the settle phase until termination.

Evaluation. On the training objects we run 800 episodes per method from randomly drawn start frames. On the held-out objects we run one episode per stable start frame and method (160 episodes).

All methods see identical start states, and policies are evaluated deterministically. Full results are in Table 11.

Table 11: Bimanual handover. *Left:* training objects, 800 episodes per method. *Right:* held-out objects, restricted to the 160 motion-capture start frames that are physically stable (Sec. E.5). All methods start from identical states.

	Training Objects			Held-Out Objects		
Method	SR ↑	TC ↑	E_t ↓	SR ↑	TC ↑	E_t ↓
Action Space Exploration	53.8	69.2	2.46	11.9	30.8	5.27
Hand-Only Prior	26.8	51.2	4.39	5.0	35.3	4.45
Ours	**69.0**	**79.4**	**1.60**	**51.9**	**74.6**	**2.18**

F Robustness and Ablations

We run all experiments in this section on Tabletop Manipulation on GRAB.

F.1 Robustness to Physical Perturbations

We test how the Tabletop Manipulation policies of all three methods cope with physical conditions that differ from training, varying one factor at a time on the held-out objects (Table 12). Changing the surface friction affects all methods similarly, and higher friction makes grasping easier for every method. Heavier objects, however, clearly separate the methods: at four times the nominal mass, our policy loses 11.7% of its success rate, while the hand-only prior and action space exploration lose 67.2% and 52.8%. Our policy is also the most robust to action noise. Under all perturbations our method retains the highest absolute success rate.

F.2 Sensitivity to the Termination Threshold

Tabletop Manipulation uses a relaxed termination threshold of 10 cm (Section 4.2). To verify that our results do not depend on this choice, we evaluate the same policies with stricter object-position thresholds of 7.5 cm and 5 cm (Table 13). All methods lose success as the threshold tightens, but our method degrades the least: from 10 cm to 5 cm, its success rate drops by 10.5 points, compared with 13.4 points for the hand-only prior and 21.9 points for action space exploration. The gap to the strongest baseline therefore grows from 14.0 to 16.9 points, which indicates that our policies track the target trajectory more closely rather than merely staying within the tolerance.

F.3 Stage III Ablations

Table 14 compares our Stage III policy with three ablated variants. The two trained variants use the same observations, reward and PPO settings as our full policy.

Direct prior rollout. This variant executes the frozen prior and decoder without any exploration policy, $\mathbf{a}_t = \mathcal{D}\big(\mathbf{o}_t^{\text{sim}}, \mu_{\mathcal{P}}(\tilde{\mathbf{o}}_t)\big)$, and thus measures how much Stage III adds on top of the distilled prior.

Action-space residual. This variant keeps the frozen prior and decoder, but the exploration policy adds its residual to the decoded action instead of to the latent, $\mathbf{a}_t = \mathcal{D}\big(\mathbf{o}_t^{\text{sim}}, \mu_{\mathcal{P}}(\tilde{\mathbf{o}}_t)\big) + \Delta\mathbf{a}_t$.

W/o random trajectories. This variant trains Stage III only on the recorded GRAB object trajectories, while keeping all other augmentations such as the randomized initial states.

F.4 Residual Bound

We ablate the residual bound B by training Stage III policies with $B \in \{0.5, 1, 2, 3, 5, 10, \infty\}$ and evaluating them on held-out objects at matched training epochs (Figure 4). The bound trades task success against grasp diversity. Tight bounds leave the policy too little room to correct the prior,

Table 12: Zero-shot robustness of the Tabletop Manipulation policies to physical perturbations. Success rate in % on held-out objects (100 evaluations per motion), varying one factor at a time; parentheses give the relative change w.r.t. each method's own base setting (friction 0.7, mass $1\times$, no noise). Action noise adds zero-mean Gaussian noise to the commanded joint targets at every control step (standard deviation in degrees for rotational joints and in mm for wrist translation). Single seed only.

(a) Surface friction

Friction	Action Space Exploration	Hand-Only Prior	Ours
0.35 (slippery)	53.8 (−8.0%)	58.5 (−6.8%)	68.5 (−5.4%)
0.7 *(base)*	58.5	62.8	72.4
1.4	63.4 (+8.4%)	67.2 (+7.0%)	76.4 (+5.5%)
2.1 (grippy)	66.1 (+13.0%)	69.0 (+9.9%)	78.2 (+8.0%)

(b) Object mass multiplier

Mass	Action Space Exploration	Hand-Only Prior	Ours
0.5× (light)	59.7 (+2.1%)	63.9 (+1.8%)	72.1 (−0.5%)
1× *(base)*	58.5	62.8	72.4
2×	48.2 (−17.6%)	40.3 (−35.8%)	71.4 (−1.4%)
4× (heavy)	27.6 (−52.8%)	20.6 (−67.2%)	64.0 (−11.7%)

(c) Action noise

Noise	Action Space Exploration	Hand-Only Prior	Ours
0 *(base)*	58.5	62.8	72.4
1° / 1 mm	54.8 (−6.3%)	52.6 (−16.2%)	67.6 (−6.6%)
2° / 2 mm	46.2 (−21.0%)	32.4 (−48.4%)	60.7 (−16.2%)
4° / 4 mm	27.0 (−53.8%)	14.6 (−76.8%)	42.6 (−41.2%)

Table 13: Tabletop Manipulation on held-out GRAB objects under stricter object-position termination thresholds. Success rate in %, mean ± standard deviation over 3 seeds.

Method	SR @ 10 cm ↑	SR @ 7.5 cm ↑	SR @ 5 cm ↑
Action Space Exploration	59.1 ± 4.7	50.0 ± 4.0	37.2 ± 4.0
Hand-Only Prior	60.0 ± 3.1	55.0 ± 2.6	46.6 ± 2.4
Ours	**74.0 ± 0.9**	**69.9 ± 1.1**	**63.5 ± 1.1**

Table 14: Stage III ablations on held-out GRAB objects. E_t in cm, E_r in degrees.

Method	SR ↑	TC ↑	E_t ↓	E_r ↓
Direct prior rollout	30.8	50.4	2.19	20.8
Action-space residual	49.6	65.2	1.87	21.0
W/o random traj.	67.9	78.4	1.37	20.4
Ours	**74.0**	**82.9**	**1.36**	**20.0**

and success rises steadily as the bound is relaxed (Figure 4a). Up to $B{=}5$, the deployed latents stay within the prior's latent distribution (Figure 4c), and hand usage closely follows the human reference (Figure 4b). For $B \geq 10$ the bound no longer constrains the residual, and $B{=}10$ behaves like the unbounded policy. Success improves only marginally, but the latents drift away from the prior and the policy collapses almost entirely to bimanual grasps, discarding the grasp diversity encoded in the prior. We therefore use $B{=}5$, which retains most of the unbounded success rate while keeping the hand usage close to the human reference.