DONIC: GENERALIST DEXTEROUS BIMANUAL TRACKING CONTROL FROM HUMAN DEMONSTRATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent progress in dexterous reference tracking, existing models focus primarily on single-hand manipulation and a limited range of interactions, leaving the bimanual dexterous behaviors common in real life unaddressed. The scarcity of contact-rich bimanual manipulation data and the increased complexity of coordinating contacts make developing a generalist bimanual tracker particularly challenging. Therefore, we present **1) DONIC**, the first generalist bimanual tracker jointly trained across operation categories and human-motion sources, together with **2) BiHOI**, a multi-view dataset of $1{,}575$ sequences spanning 229 objects and emphasizing sustained bimanual interaction, where reconstructed interactions are retargeted into robot references. DONIC combines proximal policy optimization with behavior cloning in a feedback loop that recycles verified successful executions into training demonstrations. An object-centered per-hand reward maintains object-motion feedback while supervising each hand according to its reference contact role. Clip-Centered Advantages remove reference-dependent advantage offsets while preserving within-trajectory differences. Together, these designs enable one single policy to learn coordinated manipulation across heterogeneous references. Without fine-tuning, a single checkpoint achieves 86.79% success on unseen trajectories of training objects and 57.32% on held-out objects, surpassing the state-of-the-art bimanual tracking baselines by 77.36% and 45.73%.

1 INTRODUCTION

Dexterous manipulation supports household assistance, industrial assembly, and assistive care. Unlike gripper-based manipulation, it requires precise coordination of finger motions and changing hand–object contacts to grasp, reorient, and reposition objects (Rajeswaran et al., 2018; Andrychowicz et al., 2020; Chen et al., 2023; Wan et al., 2023). Generalizing across objects and behaviors is challenging, especially when human demonstrations must be retargeted to robot hands with different kinematics (Qin et al., 2022; Liu et al., 2025; Xu et al., 2025). Simulation-ready objects (Cao et al., 2025; 2026a;b) and synthetic trajectories (Jing et al., 2025; Yin et al., 2025) enable large-scale interaction data, but may not fully capture natural human manipulation. Recent trackers therefore learn from human hand–object motion (Liu et al., 2025; Xu et al., 2025; Adalibieke et al., 2026), yet remain largely limited to single-hand tasks. Many everyday tasks instead require asymmetric bimanual coordination: one hand stabilizes an object while the other opens, rotates, or repositions it (Fan et al., 2023; Zhan et al., 2024; Zhou et al., 2026). This motivates a generalist bimanual reference tracker that learns from diverse human motion to control different objects and tasks with one policy.

Developing such a tracker poses two challenges. First, public datasets include bimanual manipulation (Fan et al., 2023; Zhan et al., 2024; Liu et al., 2024; Fu et al., 2025), but fully tracked hand–object sequences with sustained two-hand contact remain scarce, limiting the available coordination patterns. Second, a shared policy must learn trajectories with changing hand roles and contact timing. Hand-pose accuracy alone does not ensure successful object manipulation, while requiring both hands to satisfy contact conditions can mask one hand's progress. Unequal clip difficulty also creates advantage offsets that can weaken updates for useful actions in harder sequences. Shared training

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

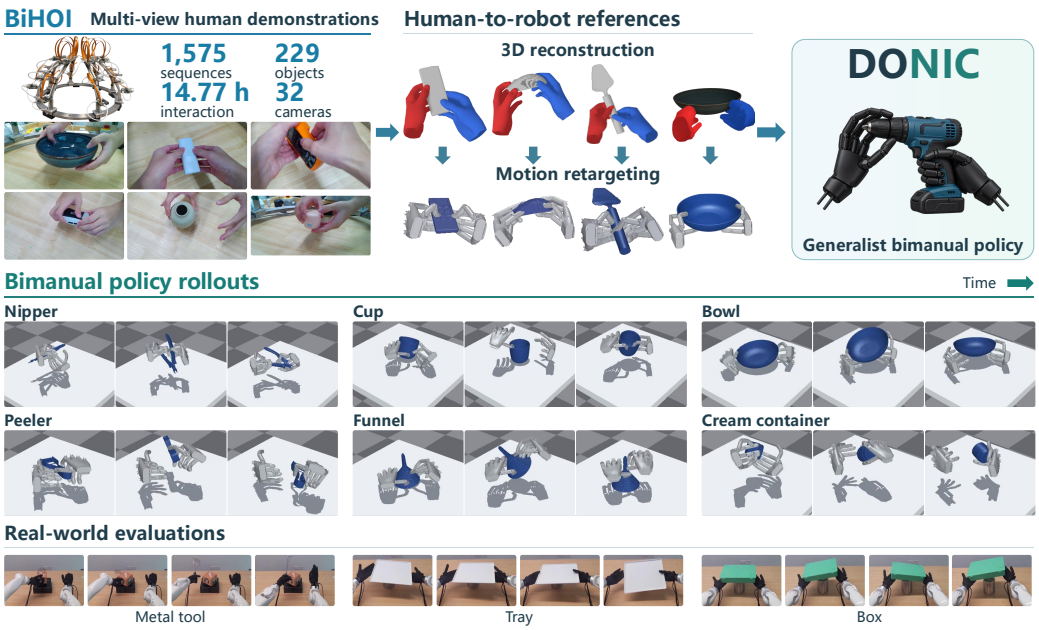

Figure 1: **Overview of DONIC.** BiHOI provides multi-view bimanual demonstrations that are reconstructed and retargeted to robot hands. DONIC learns a shared policy for diverse bimanual reference-tracking tasks.

therefore needs feedback grounded in object motion and each hand's reference contacts, alongside advantages that preserve differences among transitions within each clip.

We address these challenges with DONIC, a generalist bimanual reference tracker, and BiHOI, a multi-view hand–object interaction dataset. BiHOI contains 1,575 sequences spanning 229 objects and 14.77 hours, captured from 32 calibrated RGB views. It covers symmetric manipulation and asymmetric coordination. We reconstruct both hands and the object, then retarget hand motion to robot references paired with the object trajectory. Combined with public references, BiHOI broadens the training objects and coordination patterns. DONIC learns one policy for both hands through proximal policy optimization (PPO) (Schulman et al., 2017) and behavior cloning (BC) on verified demonstrations. Trajectory-specific policies optimize screened references, while successful shared-policy executions are verified and reused. At inference, one checkpoint tracks new references without further training. Our *object-centered per-hand reward* maintains object-motion feedback and supervises each hand's reference contacts, including asynchronous engagement. Our *Clip-Centered Advantages (CCA)* subtract each clip's mean estimated advantage before PPO updates, removing clip-dependent offsets while preserving within-clip differences. On held-out trajectories and objects from public datasets, DONIC exceeds the stronger bimanual baseline by 77.36 and 45.73 percentage points in success rate, respectively.

Our contributions are:

- To our knowledge, DONIC is the first bimanual reference tracker jointly trained across operation categories and human motion datasets; a single checkpoint is evaluated on unseen trajectories and objects without fine-tuning.
- We introduce BiHOI, a multi-view hand-object interaction dataset emphasizing coordinated bimanual manipulation, and reconstruct its recordings into robot motion references.
- We propose an object-centered per-hand reward and CCA to support joint PPO training across bimanual references with different contact roles and difficulty.
- On our two public-data evaluation splits, DONIC outperforms bimanual adaptations of DexTrack and Dexplore in success rate and completion, reaching 86.79% and 57.32% success on unseen trajectories and objects, respectively.

2 Related Work

From human references to executable manipulation. Human trajectories provide motion targets, but retargeting alone cannot ensure successful contact. Sequence-specific policies commonly use reinforcement learning to correct retargeted motion. ManipTrans (Li et al., 2025) reuses a general hand-motion imitator but learns a task-specific residual policy for interaction. DexMachina (Mandi et al., 2026) learns from individual demonstrations by gradually removing virtual object assistance. CHORD (Zhu et al., 2026) guides sequence-specific learning through contact-wrench matching, while ConTrack (Liang et al., 2026) trains one policy per clip with an adaptive balance between object tracking and motion fidelity. We instead learn one generalist bimanual policy that tracks diverse references without per-reference training or fine-tuning at test time.

Generalist manipulation policies. Generalist policies share knowledge across objects, motions, or skills, predominantly in single-hand settings. UniDexGrasp++ (Wan et al., 2023) generalizes grasping across object geometries, UniCross (Zhang et al., 2026) unifies multiple manipulation skills, and ArtManip (Yang et al., 2026) learns category-level articulated in-hand manipulation. For reference tracking, DexTrack (Liu et al., 2025) combines reinforcement learning with demonstrations mined by trajectory-specific controllers. Dexplore (Xu et al., 2025) expands exploration around reference motions, while AdaDexTrack (Adalibieke et al., 2026) combines specialist distillation with online reference modulation. Bimanual tracking must coordinate complementary hand roles and contact timing. Existing work explores narrower generalization: BiDexHD (Zhou et al., 2026) distills teachers into a separate trajectory-conditioned policy for each operation category and evaluates new object–tool combinations and instances within categories; PhysGraph (Li et al., 2026b) transfers task-trained policies to new tool and object geometries. We jointly train one bimanual reference tracker across operation categories and human motion datasets, and evaluate it on unseen trajectories and objects without fine-tuning.

Bimanual hand–object interaction datasets. GRAB (Taheri et al., 2020) records whole-body grasping with detailed contact information, while ARCTIC (Fan et al., 2023) captures the coupled motion of two hands and articulated objects. OakInk2 (Zhan et al., 2024) organizes bimanual manipulation into primitive and complex tasks, and TACO (Liu et al., 2024) covers diverse tool–action–object compositions. GigaHands (Fu et al., 2025) further expands the scale and diversity of bimanual activities with language annotations. Complementing these motion collections, H2O (Kwon et al., 2021) and HOI4D (Liu et al., 2022) support egocentric interaction understanding, while HOT3D (Banerjee et al., 2025) and HO-Cap (Wang et al., 2025a) provide recordings and annotations for hand–object reconstruction and tracking. Capture and annotation conventions differ, and human kinematics do not ensure executable robot actions. We harmonize, retarget, and screen zero-residual feasibility to form a common bimanual reference pool for joint training.

3 Bimanual Hand–Object Dataset

Table 1: Dataset overview: source statistics and simultaneous-contact rates on analyzed recordings.

Source	Capture	Sequences	Frames	Objects	Bimanual contact (%)
BiHOI (ours)	MV RGB	1,575	$\approx$1.60M†	229	**65.91**
GRAB (2020)	Marker MoCap	1,334	1.622M	51	32.21
OakInk2 (2024)	MV RGB + MoCap	627	4.01M	75	24.97
TACO (V1) (2024)	MV RGB + ego RGB-D	2.5K	5.2M	206	15.69
GigaHands (2025)	MV RGB	–	183M	417	34.74
H2O (2021)	MV RGB-D	–	571.6K	8	38.57
HOI4D (2022)	Ego RGB-D	4,000	2.4M	800	–
HO-Cap (rev.) (2025a)	MV RGB-D	64	656K	64	–

MV: multi-view; ego: egocentric. †14.77 hours at 30 fps, counting synchronized time steps once. Public frame counts follow source conventions. Contact is the total number of both-hand-contact frames divided by frames with at least one hand in contact, pooled across analyzed clips; –: unavailable.

To expand limited bimanual training data, we introduce **BiHOI**, a multi-view hand–object interaction dataset. We describe its capture system and composition, then the annotation and retargeting pipeline that converts recordings into robot references.

3.1 Multi-View Capture and Dataset

Multi-view capture system. We record hand–object interactions with 32 calibrated RGB cameras surrounding a tabletop workspace, at a resolution of 1920×1232 pixels and a frame rate of 30 fps. Complementary views support reconstruction under occlusion; calibration supplies a common metric coordinate system.

Dataset composition. BiHOI contains $1{,}575$ interaction sequences covering 229 objects and 14.77 hours of unique interaction. The objects span household items, containers, utensils, tools, and electronics. Participants perform natural object-centric interactions rather than following fixed motion templates. The recordings include symmetric bimanual manipulation, asymmetric bimanual coordination, and coordinated multi-object manipulation. Table 1 compares BiHOI with commonly used hand–object interaction datasets. Our collection emphasizes sustained two-hand interaction. Among frames with at least one hand in contact, 65.91% have both hands in contact, exceeding H2O, the highest analyzed public subset at 38.57%, by 27.34 percentage points.

3.2 Annotation and Reference Construction

Reconstruction and annotation. We reconstruct metric hand–object trajectories from synchronized RGB recordings. SAM3 provides object masks (Carion et al., 2025), while multi-view Depth Anything 3 provides spatially consistent depth that we align to metric scale using the calibrated camera extrinsics (Lin et al., 2025). MV-SAM3D reconstructs a canonical object mesh from first-frame views and masks (Li et al., 2026a); back-projected depth determines its physical scale. FoundationPose then recovers 6-DoF object motion (Wen et al., 2024), using multi-view pose consensus to handle occlusion and re-registration when consensus fails.

For each participant, triangulated calibration keypoints are used to estimate metric hand scale and calibrate a participant-specific MANO shape. WiLoR provides per-view hand observations (Potamias et al., 2025); these observations are triangulated, temporally smoothed, and fitted with the MANO hand model (Romero et al., 2017), with shape held fixed across the sequence. Joint refinement against multi-view metric depth combines geometric alignment with reconstruction and temporal regularization, producing bimanual MANO parameters, a metric object mesh, and its 6-DoF trajectory in a common world frame. Reconstruction and refinement details appear in supplementary Section B.2.

Retargeting. We map the reconstructed human motion to the Wuji robot hands through frame-wise kinematic fitting, following standard retargeting practice (Handa et al., 2020; Qin et al., 2023; Liu et al., 2025). After wrist-frame alignment, we fit hand keypoints to robot links under joint limits, initializing and regularizing each frame against the preceding solution. Smoothing occurs only before retargeting. The resulting wrist and finger references are paired with the object trajectory.

4 Methodology

DONIC learns a generalist bimanual tracker across operation categories and human-data sources (Figure 2). Its object-centered per-hand reward supports complementary roles and asynchronous contacts while maintaining object-motion feedback. CCA removes reference-dependent offsets after generalized advantage estimation (GAE) (Schulman et al., 2016), preserving within-reference advantage differences.

4.1 Task Definition

Given a reference for two hands and an object, one policy controls both robot hands to reproduce the object motion through physical interaction. Following DexTrack (Liu et al., 2025), training

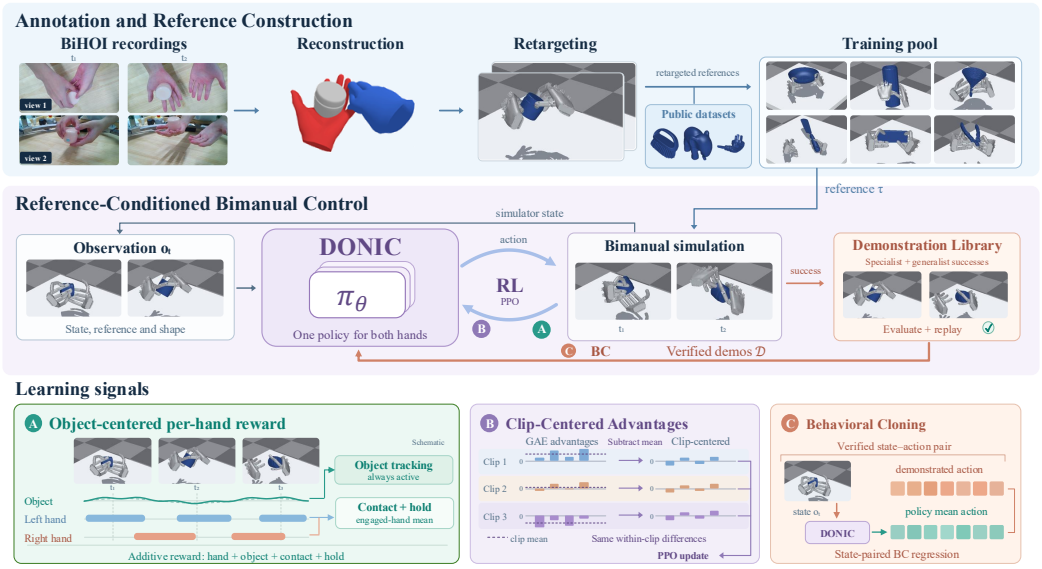

Figure 2: **Overview of DONIC. Top:** BiHOI captures are reconstructed and retargeted, then combined with public-data references to form the training pool. **Middle:** A shared bimanual policy learns through simulated interaction and verified demonstrations in an iterative PPO+BC loop. **Bottom:** An object-centered per-hand reward, Clip-Centered Advantages (CCA), and behavioral cloning provide complementary learning signals.

combines PPO with BC from executable demonstrations; the policy tracks unseen trajectories and objects without fine-tuning.

For a clip spanning T control steps, we write the reference as

$$\tau = \left\{(\hat{\boldsymbol{q}}_t^L, \hat{\boldsymbol{q}}_t^R, \hat{\boldsymbol{p}}_t^o, \hat{\boldsymbol{R}}_t^o)\right\}_{t=0}^{T}. \tag{1}$$

Here t is the frame index, $h \in \{L, R\}$ identifies a hand, and a hat denotes a reference quantity. The configuration $\hat{\boldsymbol{q}}_t^h \in \mathbb{R}^d$ contains the wrist coordinates and finger-joint angles, with d controlled coordinates per hand. The object position and orientation are $\hat{\boldsymbol{p}}_t^o \in \mathbb{R}^3$ and $\hat{\boldsymbol{R}}_t^o \in SO(3)$, where $SO(3)$ is the space of three-dimensional rotations. We concatenate the two hand configurations into $\hat{\boldsymbol{q}}_t \in \mathbb{R}^{2d}$. Corresponding quantities without hats denote simulated states. Object linear velocity is denoted by $\boldsymbol{v}_t^o$; its reference $\hat{\boldsymbol{v}}_t^o$ is obtained by finite differences. Positions, rotations, and linear velocities use metres, radians, and metres per second, respectively.

4.2 Reference-Conditioned Bimanual Control

A single stochastic policy $\pi_\theta(\boldsymbol{a}_t \mid \boldsymbol{o}_t)$, parameterized by θ, predicts joint actions $\boldsymbol{a}_t$ for both hands from observation $\boldsymbol{o}_t$. Following reference-based trackers (Liu et al., 2025; Xu et al., 2025), $\boldsymbol{o}_t$ concatenates two per-hand blocks containing proprioception, object pose and velocity, upcoming reference targets, reference-state errors, and fingertip-to-object distance vectors. Spatial features use each wrist's heading frame. We append a basis-point-set shape encoding (Prokudin et al., 2019), object scale and centroid, hand identities, and the fixed left-hand reward weight $w_L \in (0, 1]$. Contact-force channels are zeroed; simulated forces supervise rewards only.

Given $\boldsymbol{o}_t$, the policy predicts bounded corrections $\boldsymbol{a}_t \in [-1, 1]^{2d}$ to the next reference, producing a proportional–derivative (PD) control target as in prior trackers (Liu et al., 2025; Mandi et al., 2026):

$$\boldsymbol{q}_t^{\mathrm{pd}} = \mathrm{clamp}(\hat{\boldsymbol{q}}_{t+1} + \boldsymbol{s} \odot \boldsymbol{a}_t, \boldsymbol{q}_{\min}, \boldsymbol{q}_{\max}), \tag{2}$$

where $\boldsymbol{s} \in \mathbb{R}^{2d}$ sets the correction range, $\odot$ denotes element-wise multiplication, and clamp enforces the coordinate limits $\boldsymbol{q}_{\min}$ and $\boldsymbol{q}_{\max}$. Corrections are anchored to the next reference at every step and do not accumulate. The PD controller executes the target, and the resulting state supplies

the next observation. We build on this control interface with the following reward and training procedure.

4.3 Object-Centered Per-Hand Reward

Hand-pose accuracy does not ensure contact or object motion, and the hands may engage asynchronously with different roles. We therefore reward object motion throughout the clip while supervising each hand's contact only when required by the reference.

Tracking the object. We reward agreement in object position, orientation, and linear velocity:

$$r_t^{\text{obj}} = e^{-\kappa_p \|\boldsymbol{p}_t^o - \hat{\boldsymbol{p}}_t^o\|_2} e^{-\kappa_R \angle(\boldsymbol{R}_t^o, \hat{\boldsymbol{R}}_t^o)} e^{-\kappa_v \|\boldsymbol{v}_t^o - \hat{\boldsymbol{v}}_t^o\|_2}. \tag{3}$$

Positive coefficients $\kappa_p, \kappa_R, \kappa_v$ control sensitivity to position, rotation, and velocity errors. Here $\|\cdot\|_2$ is the Euclidean norm and $\angle(\cdot,\cdot)$ the rotation geodesic distance. The product rewards joint agreement, with velocity penalizing stationary solutions to moving references. This kinematic term (Wang et al., 2023; 2025b) is independent of the contact gate.

Supervising each hand's contact. Object tracking alone does not specify how the hands should support the motion. Let $\mathcal{C}_t^h$ be the set of fingertips on hand h required to contact the object by the reference at frame t. For fingertip i, let $\delta_{i,t}^h$ denote its distance to the object surface and $f_{i,t}^h$ its simulated net contact-force magnitude. We estimate contact using

$$\alpha_{i,t}^h = \mathbf{1}[f_{i,t}^h > f_{\min}]\, \mathbf{1}[\delta_{i,t}^h < \delta_{\max}], \tag{4}$$

where $\mathbf{1}[\cdot]$ is an indicator and $f_{\min}, \delta_{\max} > 0$ are force and distance thresholds. Proximity reduces false object-contact attribution, but net force does not identify the collision partner.

For a hand with at least one required fingertip, we define

$$\begin{aligned} r_t^{\text{cont},h} &= \frac{1}{|\mathcal{C}_t^h|} \sum_{i \in \mathcal{C}_t^h} \left[\zeta e^{-\kappa_c \delta_{i,t}^h} + (1-\zeta)\alpha_{i,t}^h \right], \\ r_t^{\text{hold},h} &= \frac{1}{|\mathcal{C}_t^h|} \sum_{i \in \mathcal{C}_t^h} \alpha_{i,t}^h. \end{aligned} \tag{5}$$

Here $|\mathcal{C}_t^h|$ is the number of required fingertips, $\kappa_c > 0$ controls distance sensitivity, and $\zeta \in [0,1]$ balances proximity and contact. Proximity guides contact acquisition; the hold term rewards instantaneous required-contact coverage, allowing valid sliding or rolling.

Combining the learning signals. We aggregate contact rewards over the engaged-hand set $\mathcal{E}_t = \{h \in \{L,R\} : \mathcal{C}_t^h \neq \varnothing\}$. For $u \in \{\text{cont}, \text{hold}\}$, let

$$r_t^u = \begin{cases} \dfrac{\sum_{h \in \mathcal{E}_t} w_h r_t^{u,h}}{\sum_{h \in \mathcal{E}_t} w_h}, & \mathcal{E}_t \neq \varnothing, \\ 0, & \mathcal{E}_t = \varnothing, \end{cases} \tag{6}$$

where $w_R = 1$ and $w_L \in (0,1]$ are fixed hand weights. Averaging preserves either engaged hand's feedback when the other loses contact; excluding inactive hands avoids rewarding attachment during reference release.

Finally, we combine these signals with the inherited hand-tracking reward $r_t^{\text{hand}} \in [0,1]$, which measures wrist, finger-joint, and keypoint agreement (Xu et al., 2025):

$$r_t = (1 + w_L)\left[\beta_{\text{hand}} r_t^{\text{hand}} + \beta_{\text{obj}} r_t^{\text{obj}} + \beta_{\text{cont}} r_t^{\text{cont}} + \beta_{\text{hold}} r_t^{\text{hold}}\right]. \tag{7}$$

The nonnegative β coefficients sum to one; $(1 + w_L)$ sets the reward scale. These terms jointly supervise hand motion, object outcomes, and reference contact roles (settings in supplementary Section C.2).

4.4 Learning a Generalist from Verified Demonstrations

Following DexTrack (Liu et al., 2025), we alternate PPO+BC training with demonstration collection. PPO learns through simulated interaction, while BC supplies successful actions that shared exploration may struggle to discover. Trajectory-specific policies optimize references retained by replay screening; their verified rollouts form a demonstration library $\mathcal{D}$ of observation–action pairs. As the generalist improves, its verified successes are added to $\mathcal{D}$ for subsequent updates. All demonstrations come from training references.

Demonstrations supply successful behavior, while CCA addresses unequal reference difficulty in PPO updates.

Clip-Centered Advantages. Shared PPO batches mix contact patterns and coordination demands. GAE combines temporally weighted value-prediction residuals, but unequal critic accuracy across references can leave clip-dependent offsets. Difficult clips may then receive consistently negative advantages despite relatively successful transitions. CCA removes this common offset while preserving within-clip differences.

Formally, let $\mathcal{B}$ denote an on-policy rollout batch, j index a transition, and $m(j)$ identify its reference clip. After computing the standard GAE advantage A_j^{GAE} from rewards and the learned value function, we group transitions by reference as $\mathcal{G}_m = \{j \in \mathcal{B} : m(j) = m\}$ and use

$$\widetilde{A}_j = A_j^{\mathrm{GAE}} - \frac{1}{|\mathcal{G}_{m(j)}|} \sum_{k \in \mathcal{G}_{m(j)}} A_k^{\mathrm{GAE}}. \tag{8}$$

Subtracting the clip mean preserves pairwise advantage differences among its transitions while removing their common empirical offset. The adjustment is applied after GAE; the learned critic and temporal advantage estimation remain unchanged.

Groups are defined by reference identity rather than by environment instance or temporal phase, allowing parallel rollouts of the same reference to contribute to a common baseline. Unlike batch-wide centering, CCA removes an offset specific to each reference instead of applying one shared shift to all clips. We subtract only the group mean, without dividing by a per-clip standard deviation, preserving the scale of within-clip advantage differences. This adjustment acts on the policy's learning signal; it does not modify the reward or the critic's regression targets.

Joint PPO and BC updates. We use $\widetilde{A}_j$ in the standard PPO actor loss $\mathcal{L}_{\mathrm{PPO}}$, retaining its on-policy probability ratio and clipping rule. For a demonstration observation–action pair $(\boldsymbol{o}^D, \boldsymbol{a}^D) \in \mathcal{D}$, let $\mu_\theta(\boldsymbol{o}^D)$ denote the shared Gaussian policy's mean action. The combined actor objective is

$$\mathcal{L}_{\mathrm{actor}}(\theta) = \mathcal{L}_{\mathrm{PPO}}(\theta; \widetilde{A}) + \lambda \mathbb{E}_{(\boldsymbol{o}^D, \boldsymbol{a}^D) \sim \mathcal{D}} \left\| \mu_\theta(\boldsymbol{o}^D) - \boldsymbol{a}^D \right\|_2^2, \tag{9}$$

where $\lambda > 0$ weights BC and the expectation averages over sampled demonstration pairs. Demonstrations use a separate sampling budget and do not enter the PPO ratio or GAE. The critic retains its standard value-regression objective. Sampling, verification, and training details are provided in Sections C.4–C.5 of the supplementary material.

5 Experiments

5.1 Setup

Datasets. We curate bimanual manipulation references from GRAB (Taheri et al., 2020), GigaHands (Fu et al., 2025), OakInk2 (Zhan et al., 2024), TACO (Liu et al., 2024), and H2O (Kwon et al., 2021). We retain clips in which both hands and the manipulated object are tracked, and each hand contacts the object at some point during the clip; simultaneous contact is not required. Motions are harmonized, retargeted to the Wuji Hand V1, and replay-screened (supplementary Section B.1). This yields $1{,}845$ public-data training clips, separate from the held-out evaluation references. Ours+BiHOI additionally uses all $1{,}575$ BiHOI sequences, bringing its training pool to $3{,}420$ clips.

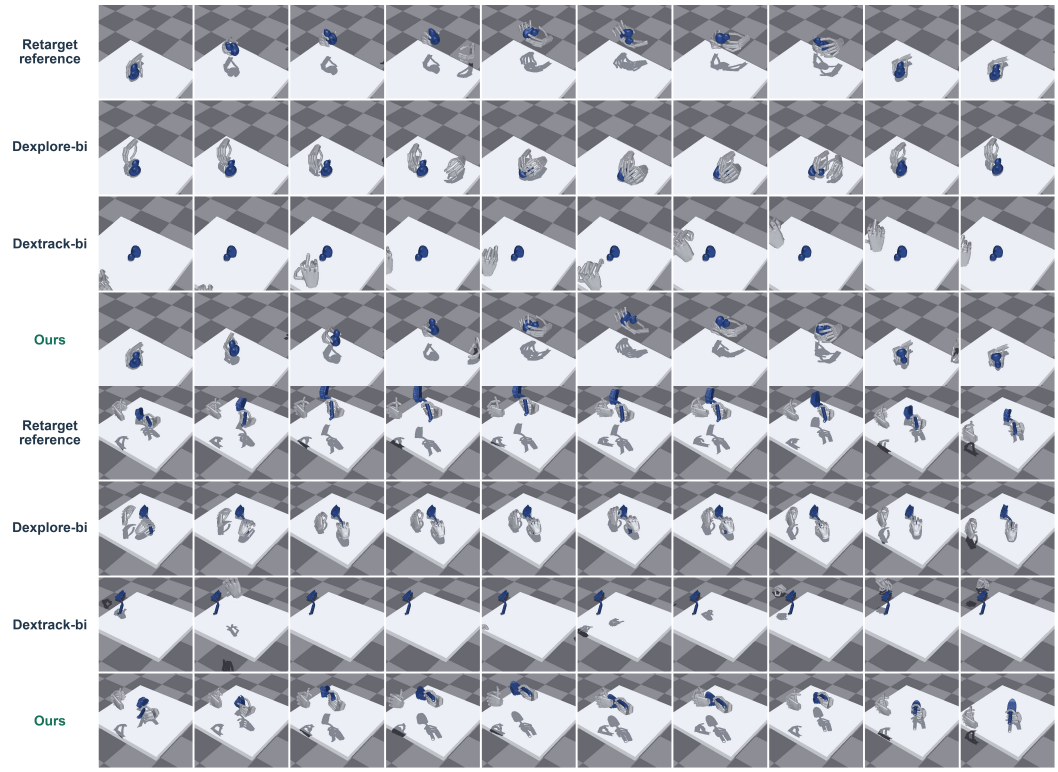

Figure 3: Qualitative comparisons on two held-out object trajectories. Rows show the reference, Dexplore-bi, DexTrack-bi, and DONIC; time runs left to right.

For evaluation, we reserve two splits from the public datasets: UNSEEN-TRAJ contains 53 unseen trajectories of training objects, and UNSEEN-OBJ contains 164 clips involving held-out objects. Neither split is used for policy training or demonstration collection. All methods are evaluated on the same splits without fine-tuning.

Metrics. We extend AdaDexTrack's evaluation protocol (Adalibieke et al., 2026) to bimanual tracking. We report object errors (T_o, R_o), wrist errors (T_h, R_h), and finger-joint error (E_{finger}), with translation in centimetres and angles in radians. Wrist errors use robot forward kinematics against retargeted references; finger error averages over 20 finger joints. Reported hand errors take the worse of the two trajectory-averaged hand errors, then average across trajectories. SR requires both hands independently to satisfy $\frac{1}{2}(T_o + T_h) \leq 10\,\text{cm}$, $\frac{1}{2}(R_o + R_h) \leq \pi/3$, and $E_{\text{finger}} \leq 2\pi/9$ on trajectory-averaged errors. Comp is the fraction of frames satisfying all three conditions for both hands, averaged over trajectories.

Baselines. We compare against *DexTrack-bi* and *Dexplore-bi*, using their released code with observation and control interfaces adapted to two Wuji hands. Both baselines share DONIC's public-data pool, Isaac Gym simulator, robot model, and held-out splits. DexTrack-bi follows the released specialist–distillation pipeline (Liu et al., 2025). Dexplore-bi retains Dexplore's original reward and reference-scoped exploration (Xu et al., 2025). Additional implementation details are provided in supplementary Section D.3.

5.2 Main Results

One DONIC checkpoint improves both held-out splits without fine-tuning (Table 2). On UNSEEN-TRAJ, DONIC achieves 86.79% SR and 67.62% Comp, exceeding the stronger baseline, Dexplore-bi, by 77.36 and 39.75 percentage points, respectively. On UNSEEN-OBJ, it reaches 57.32% SR and 54.69% Comp, compared with 11.59% and 26.37% for Dexplore-bi. The gains extend to hand

Table 2: Reference tracking without fine-tuning, using one checkpoint per method across both splits. Black and blue bold mark the best and second-best results within each split; ties share highlighting.

Dataset	Method	$T_o \downarrow$ (cm)	$R_o \downarrow$ (rad)	$T_h \downarrow$ (cm)	$R_h \downarrow$ (rad)	$E_{\text{finger}} \downarrow$ (rad)	SR↑ (%)	Comp↑ (%)
Unseen trajectory	DexTrack-bi	41.86	2.3340	23.03	0.8296	0.6301	0.00	1.30
	Dexplore-bi	14.73	**0.8597**	21.88	0.8412	0.3166	9.43	27.87
	w/o CCA	11.27	0.9585	4.34	0.2980	0.0433	81.13	62.72
	w/o contact gate	**10.26**	0.9589	4.00	0.3013	**0.0338**	81.13	65.03
	Ours	10.27	0.9277	**3.96**	**0.2905**	**0.0372**	**86.79**	**67.62**
	Ours+BiHOI	**9.74**	**0.8637**	**3.32**	**0.2818**	**0.0338**	**88.68**	**68.72**
Unseen object	DexTrack-bi	127.87	2.3532	72.83	1.0629	0.6455	0.00	0.88
	Dexplore-bi	17.92	1.1431	19.46	0.9509	0.3685	11.59	26.37
	w/o CCA	16.83	1.1464	5.52	0.3235	0.0378	48.17	43.89
	w/o contact gate	17.61	1.2041	4.97	0.3386	**0.0371**	48.78	49.52
	Ours	**15.74**	**1.1310**	**4.58**	**0.3226**	0.0373	**57.32**	**54.69**
	Ours+BiHOI	**13.96**	**1.1157**	**4.54**	**0.3206**	**0.0361**	**59.76**	**55.35**

tracking: wrist translation error decreases from 21.88 to 3.96 cm on unseen trajectories and from 19.46 to 4.58 cm on unseen objects. DONIC reduces all reported errors relative to both baselines except object rotation on UNSEEN-TRAJ. Because SR and Comp require both hands to meet the tracking criteria, these improvements reflect successful bimanual execution rather than accuracy from only one hand.

On the two held-out object trajectories in Figure 3, DONIC coordinates both hands and more closely reproduces reference object motion. Dexplore-bi moves its hands but produces limited object motion, whereas DexTrack-bi loses the hand–object alignment needed to execute the reference.

5.3 Ablation Studies

Learning signals. Removing CCA reduces SR from 86.79% to 81.13% on unseen trajectories and from 57.32% to 48.17% on unseen objects, while lowering Comp and increasing all five tracking errors (Table 2). Removing the contact-ownership gate reduces SR to 81.13% and 48.78%, respectively. Despite lower individual errors in some cases, this variant also has lower Comp on both splits. Both learning signals thus improve joint bimanual success and completion.

Additional bimanual data. Adding all 1,575 BiHOI sequences expands training from 1,845 to 3,420 clips and improves all seven metrics on both splits (Ours+BiHOI, Table 2). SR reaches 88.68% and 59.76% on unseen trajectories and objects, with Comp of 68.72% and 55.35%. Unseen-object translation error falls by 11.3%, from 15.74 to 13.96 cm. Ranking first on all unseen-object metrics, Ours+BiHOI demonstrates the benefit of additional bimanual data.

6 Conclusion, Limitations and Future Work.

We presented DONIC, a generalist bimanual tracker, and BiHOI, a multi-view bimanual interaction dataset. Object-centered per-hand rewards and CCA support iterative PPO+BC training, yielding one checkpoint that generalizes to unseen trajectories and objects without fine-tuning and substantially outperforms adapted baselines. Ablations support both learning signals and the benefit of additional BiHOI data. However, our quantitative evaluation focuses on simulated single-object interactions and relies on reconstructed, retargeted motion references. We aim to extend DONIC to longer-horizon and multi-object manipulation, evaluate on an independent benchmark, and investigate transfer to physical bimanual systems.

AI USE STATEMENT

Generative AI tools assisted with language polishing to improve the clarity and readability of the manuscript. The authors are responsible for the final content.

ETHICS STATEMENT

BiHOI records human hand–object interactions in a multi-view tabletop setting. All participants consented to data collection and public release. The use and redistribution of third-party datasets remain subject to their original licenses and access conditions.

REPRODUCIBILITY STATEMENT

Dataset construction, annotation, and retargeting are detailed in Appendix B. Appendix C specifies the controller, learning signals, demonstration collection, and training configurations. Appendix D describes the evaluation splits, bimanual metrics, baseline implementations, and additional comparisons.

REFERENCES

Jianibieke Adalibieke, Qianwei Han, Xueyi Liu, Yuzhe Qin, and Li Yi. AdaDexTrack: Dynamic modulation for adaptive and generalizable dexterous manipulation tracking. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. URL https://janebek.github.io/AdaDexTrack/.

Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020. URL https://arxiv.org/abs/1808.00177.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7071, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Banerjee_HOT3D_Hand_and_Object_Tracking_in_3D_from_Egocentric_Multi-View_CVPR_2025_paper.html.

Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: Physical-grounded 3d asset generation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 104070–104083, 2025.

Ziang Cao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-anything: Simulation-ready physical 3d assets from single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5839–5848, 2026a.

Ziang Cao, Yinghao Liu, Haitian Li, Runmao Yao, Fangzhou Hong, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-omni: Unified simulation-ready physical 3d generation for rigid, deformable, and articulated objects. *arXiv preprint arXiv:2605.21572*, 2026b.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand reorientation of novel and complex object shapes. *Science Robotics*, 8(84), 2023. URL https://arxiv.org/abs/2211.11744.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fan_ARCTIC_A_Dataset_for_Dexterous_Bimanual_Hand-Object_Manipulation_CVPR_2023_paper.html.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. GigaHands: A massive annotated dataset of bimanual hand activities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17461–17474, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_GigaHands_A_Massive_Annotated_Dataset_of_Bimanual_Hand_Activities_CVPR_2025_paper.html.

Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. DexPilot: Vision-based teleoperation of dexterous robotic hand-arm system. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

Zhi Jing, Siyuan Yang, Jicong Ao, Ting Xiao, Yu-Gang Jiang, and Chenjia Bai. Humanoidgen: Data generation for bimanual dexterous manipulation via llm reasoning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 173413–173459, 2025.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Baicheng Li, Dong Wu, Jun Li, Shunkai Zhou, Zecui Zeng, Lusong Li, and Hongbin Zha. MV-SAM3D: Adaptive multi-view fusion for layout-aware 3D generation. *arXiv preprint arXiv:2603.11633*, 2026a. URL https://arxiv.org/abs/2603.11633.

Kailin Li, Puhao Li, Tengyu Liu, Yuyang Li, and Siyuan Huang. ManipTrans: Efficient dexterous bimanual manipulation transfer via residual learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://arxiv.org/abs/2503.21860.

Runfa Blark Li, David Kim, Xinshuang Liu, Keito Suzuki, Dwait Bhatt, Nikola Raicevic, Xin Lin, Ki Myung Brian Lee, Nikolay Atanasov, and Truong Nguyen. PhysGraph: Physically-Grounded Graph-Transformer Policies for Bimanual Dexterous Hand–Tool–Object Manipulation. *arXiv preprint arXiv:2603.01436*, 2026b. URL https://arxiv.org/abs/2603.01436.

Yutong Liang, Quanquan Peng, Ri-Zhao Qiu, and Xiaolong Wang. ConTrack: Constrained hand motion tracking with adaptive trade-off control. *arXiv preprint arXiv:2606.03177*, 2026. URL https://arxiv.org/abs/2606.03177.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. URL https://arxiv.org/abs/2511.10647.

Xueyi Liu, Jianibieke Adalibieke, Qianwei Han, Yuzhe Qin, and Li Yi. DexTrack: Towards generalizable neural tracking control for dexterous manipulation from human references. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2502.09614.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. TACO: Benchmarking generalizable bimanual tool-ACtion-Object understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_TACO_Benchmarking_Generalizable_Bimanual_Tool-ACtion-Object_Understanding_CVPR_2024_paper.html.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4D egocentric dataset for category-level human-object interaction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

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. *arXiv preprint arXiv:2108.10470*, 2021.

Zhao Mandi, Yifan Hou, Dieter Fox, Yashraj Narang, Ajay Mandlekar, and Shuran Song. DexMachina: Functional retargeting for bimanual dexterous manipulation. In *International Conference on Machine Learning (ICML)*, 2026. URL https://project-dexmachina.github.io/.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, 2025. doi: 10.1109/CVPR52734.2025.01143.

Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Yuzhe Qin, Yueh-Hua Wu, Shaowei Liu, Hanwen Jiang, Ruihan Yang, Yang Fu, and Xiaolong Wang. DexMV: Imitation learning for dexterous manipulation from human videos. In *European Conference on Computer Vision (ECCV)*, 2022. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136990562.pdf.

Yuzhe Qin, Wei Yang, Binghao Huang, Karl Van Wyk, Hao Su, Xiaolong Wang, Yu-Wei Chao, and Dieter Fox. AnyTeleop: A general vision-based dexterous robot arm-hand teleoperation system. In *Robotics: Science and Systems (RSS)*, 2023.

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. In *Robotics: Science and Systems (RSS)*, 2018. URL https://arxiv.org/abs/1709.10087.

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

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

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

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.

Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. UniDexGrasp++: Improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3891–3902, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Wan_UniDexGrasp_Improving_Dexterous_Grasping_Policy_Learning_via_Geometry-Aware_Curriculum_and_ICCV_2023_paper.html.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-Cap: A capture system and dataset for 3D reconstruction and pose tracking of hand-object interaction. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-5683.

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

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b. URL https://arxiv.org/abs/2408.15270.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D pose estimation and tracking of novel objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879, 2024. doi: 10.1109/CVPR52733.2024.01692.

Sirui Xu, Yu-Wei Chao, Liuyu Bian, Arsalan Mousavian, Yu-Xiong Wang, Liang-Yan Gui, and Wei Yang. Dexplore: Scalable neural control for dexterous manipulation from reference-scoped exploration. In *Conference on Robot Learning (CoRL)*, 2025. URL https://arxiv.org/abs/2509.09671.

Yang Yang, Tengyu Liu, Puhao Li, Zeyuan Chen, Yuyang Li, Xingwan Wang, Yingying Wu, Zhaopeng Cui, and Siyuan Huang. ArtManip: Category-Level Articulated In-Hand Manipulation. *arXiv preprint arXiv:2609.12498*, 2026. URL https://arxiv.org/abs/2609.12498.

Zhao-Heng Yin, Changhao Wang, Luis Pineda, Francois Hogan, Krishna Bodduluri, Akash Sharma, Patrick Lancaster, Ishita Prasad, Mrinal Kalakrishnan, Jitendra Malik, et al. Dexteritygen: Foundation controller for unprecedented dexterity. *arXiv preprint arXiv:2502.04307*, 2025.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OakInk2: A dataset of bimanual hands-object manipulation in complex task completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Zhan_OAKINK2_A_Dataset_of_Bimanual_Hands-Object_Manipulation_in_Complex_Task_CVPR_2024_paper.html.

Hui Zhang, Julian Ferchow, Jie Song, and Mirko Meboldt. UniCross: Unified cross-skill dexterous manipulation synthesis. *arXiv preprint arXiv:2607.28198*, 2026. URL https://arxiv.org/abs/2607.28198.

Bohan Zhou, Haoqi Yuan, Yuhui Fu, and Zongqing Lu. Learning diverse bimanual dexterous manipulation skills from human demonstrations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://doi.org/10.1609/aaai.v40i34.40127.

Xinghao Zhu, Zixi Liu, Shalin Jain, Chenran Li, Milad Noori, Michael Andres Lin, Huihua Zhao, John Welsh, Mrinal Verghese, Wei Liu, Tingwu Wang, Xingye Da, Zhengyi Luo, Vishal Kulkarni, Naema Bhatti, Yuke Zhu, Linxi Fan, Bowen Wen, Danfei Xu, Soha Pouya, and Yan Chang. Learning dexterous manipulation using contact wrench guidance from human demonstration. *arXiv preprint arXiv:2607.00033*, 2026. URL https://arxiv.org/abs/2607.00033.

SUPPLEMENTARY MATERIAL

This appendix provides additional related work (Appendix A), dataset construction and retargeting details (Appendix B), controller and training specifications (Appendix C), and experimental protocols and results (Appendix D).

A ADDITIONAL RELATED WORK

Executable human references and shared control. Human demonstrations provide motion targets but do not directly specify robot actions: differences in hand kinematics and simulated contact can make a kinematically plausible reference physically difficult to execute. Reference-based trackers address this gap through closed-loop control. DexTrack (Liu et al., 2025) combines trajectory-specific demonstration generation with generalist learning; Dexplore (Xu et al., 2025) expands exploration around reference motions; and AdaDexTrack (Adalibieke et al., 2026) combines specialist distillation with reference modulation. Our reference-conditioned action interface and iterative reinforcement–imitation learning loop build on this line of work. Our contributions within that loop are the object-centered per-hand reward and Clip-Centered Advantages (CCA), designed for shared bimanual tracking across heterogeneous references.

What generalization means here. Generalization across grasped objects, manipulation skills, and complete reference trajectories imposes different requirements. UniDexGrasp++ (Wan et al., 2023) studies generalizable dexterous grasping, while BiDexHD (Zhou et al., 2026) learns trajectory-conditioned bimanual policies within operation categories. Our controller is jointly trained across operation categories and human-motion sources. We evaluate new trajectories of training objects and trajectories involving held-out objects with the same checkpoint. These two splits test reference and object generalization; they do not, by themselves, establish generalization to held-out operation categories or arbitrary manipulation goals.

Interaction capture and usable control data. GRAB, GigaHands, OakInk2, TACO, and H2O provide complementary sources of hand–object interaction (Taheri et al., 2020; Fu et al., 2025; Zhan et al., 2024; Liu et al., 2024; Kwon et al., 2021). HOI4D and HO-Cap further broaden capture settings and annotation tools (Liu et al., 2022; Wang et al., 2025a). Our comparison distinguishes the inventory of a source dataset from the subset retained for robot training. A clip enters our bimanual pool only if the required hand and object tracks are available and both hands contact the object at some point. This allows sequential as well as simultaneous hand participation. BiHOI complements these sources with a higher measured proportion of simultaneous bimanual contact in the analyzed recordings.

B DATASET AND REFERENCE CONSTRUCTION

B.1 CAPTURE, COMPOSITION, AND TRAINING INVENTORY

BiHOI contains $1{,}575$ interaction sequences covering 229 objects and 14.77 hours of unique interaction. The objects span household items, containers, utensils, tools, and electronics. A calibrated array of 32 RGB cameras records a tabletop workspace at 1920×1232 pixels and 30 frames per second. Duration counts unique interaction time, not the sum over camera views. Recordings include symmetric manipulation, asymmetric coordination, and coordinated multi-object interactions. Participants perform natural object-centric interactions rather than fixed motion templates. These describe the capture inventory; the control formulation tracks two robot hands and one reference object.

We characterize contact as simultaneous or sequential and measure the duration of simultaneous contact. In the analyzed BiHOI recordings, simultaneous bimanual contact accounts for 65.91% of contact-active time (at least one hand in contact). The highest proportion among the analyzed public subsets is 38.57% for H2O, a difference of 27.34 percentage points. Table S1 reports the selected public training inventory and the measured contact proportions. These subset measurements should not be interpreted as statistics over every recording released by each source.

Table S1: Training inventory and simultaneous bimanual contact. Contact proportions pool frames across analyzed recordings, dividing both-hand-contact frames by frames with at least one hand in contact. The public-only model uses the public subsets listed below; Ours+BiHOI adds all BiHOI sequences.

Source	Training sequences	Simultaneous contact (%)
GRAB	505	32.21
GigaHands	592	34.74
OakInk2	449	24.97
H2O	127	38.57
TACO	172	15.69
Public-data pool	1,845	—
BiHOI	1,575	65.91
Public-data pool + BiHOI	3,420	—

For the public pool, both hands and the manipulated object must be tracked, and each hand must contact the object during the clip; continuous or simultaneous contact is not an admission requirement. Motions are converted to a common representation consisting of per-hand MANO parameters, an object mesh, and object position and orientation in a gravity-aligned world frame. Retargeting and replay screening then produce robot references. The 53 unseen-trajectory and 164 unseen-object evaluation clips are separate from the training counts.

Contact measurement. For clip m, frame t, and hand h, let $c^h_{m,t,i}$ denote the reference label at fingertip slot $i \in \{3, 6, 9, 12, 15\}$. Define activity $a^h_{m,t} = \mathbf{1}[\exists i : c^h_{m,t,i} > 0.01]$, using the same reference-hand activity rule as training. The reported percentage is

$$100 \frac{\sum_{m,t} a^L_{m,t} a^R_{m,t}}{\sum_{m,t} \mathbf{1}[a^L_{m,t} + a^R_{m,t} > 0]}. \tag{B.1}$$

Thus the denominator is contact-active time, and aggregation pools frames across clips. It is neither the fraction of all sequence frames nor a mean of per-clip percentages. On the analyzed public pool, this measure is 31.40%; using all frames as the denominator instead gives 22.56%.

Single-object representation. For multi-object recordings, conversion retains one primary object: TACO keeps the tool rather than the target, and OakInk2 selects the main object from its hand-role program annotations. The controller represents the selected object by a fixed mesh and a rigid pose. Articulation and deformation are not modeled explicitly, even when the captured object belongs to those categories.

B.2 Multi-View Reconstruction and Annotation

The reconstruction pipeline uses learned image predictions to initialize geometry, then calibrated multi-view observations to recover physical scale and temporal consistency. Its five stages are mask/depth extraction, object reconstruction and tracking, participant-specific hand calibration, hand-motion reconstruction, and joint refinement. Figure S1 summarizes the complete calibrated multi-view reconstruction and annotation pipeline.

Masks and metric depth. We use SAM3 (Carion et al., 2025) in two stages. The first frames from all cameras form an ordered cross-camera sequence: a single object prompt initializes consistent masks across views. Each camera video is then tracked temporally using its first-frame mask. At each synchronized time, Depth Anything 3 (Lin et al., 2025) processes the views with calibrated intrinsics and world-to-camera extrinsics. Extrinsic-scale alignment anchors depth to metric camera translations. The resulting depth supports object-scale recovery and subsequent pose refinement.

Metric object reconstruction. MV-SAM3D (Li et al., 2026a) reconstructs a canonical mesh from distributed first-frame views and masks. We back-project masked depth from valid views into the

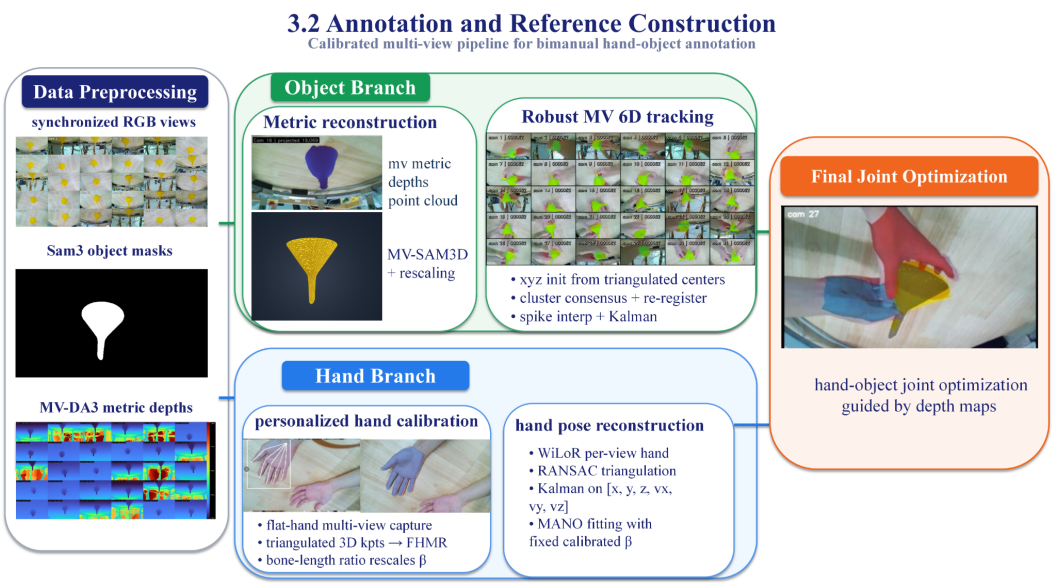

Figure S1: **Reconstruction and annotation pipeline.** Synchronized multi-view RGB observations are processed to obtain SAM3 object masks and metric depth maps. The object branch performs metric reconstruction followed by robust multi-view 6D tracking, while the hand branch performs participant-specific hand calibration and multi-view hand-pose reconstruction. The two branches are finally refined jointly using depth-guided hand–object optimization.

common world frame, reject inconsistent per-camera point-cloud centers, and fuse the remaining observations. To estimate physical scale without aligning an ambiguous canonical orientation, we compare robust principal-component extents of the depth cloud and mesh samples. After removing radial outliers, let e_k^{depth} and e_k^{mesh} be their sorted 1–99 percentile extents along the three principal axes. The isotropic scale is

$$s_{\text{obj}} = \text{median}_{k\in\{1,2,3\}} \frac{e_k^{\text{depth}}}{e_k^{\text{mesh}}}. \tag{B.2}$$

Only scale is transferred; the canonical mesh orientation is retained.

Multi-view object tracking. We triangulate centers of minimum-area rectangles fitted to object masks to initialize translation, and initialize orientation from the preceding accepted pose. The ten views with the largest visible masks are refined independently with FoundationPose (Wen et al., 2024). Candidate poses are transformed into the world frame and grouped by geometric agreement. Two candidates $(R_i, \mathbf{t}_i)$ and $(R_j, \mathbf{t}_j)$ are compatible when their rotation geodesic distance $d_{\text{SO(3)}}(R_i, R_j) < 30^\circ$ and $\|\mathbf{t}_i - \mathbf{t}_j\|_2 < 5\,\text{cm}$. At least four mutually compatible views are required. Translation is averaged with mask-area weights, while orientation is taken from the largest-mask view in the accepted cluster. Failed consensus triggers re-registration.

The object trajectory undergoes translation spike repair and Kalman filtering, followed by rotation spike repair and filtering on $SO(3)$, the space of three-dimensional rotations. Interpolation is restricted to short spike bursts. We subsequently refine the full trajectory against fused metric depth using alternating iterative closest point correspondence updates. Correspondences farther than 3 cm are rejected. For point-to-plane and point-to-point residuals r_{p2pl} and r_{p2p}, respectively, the geometric contribution is

$$\mathcal{L}_{\text{geom}} = 0.7\,\rho(r_{\text{p2pl}}) + 0.3\,\rho(r_{\text{p2p}}), \tag{B.3}$$

where ρ is the robust residual penalty. Reconstruction priors and first- and second-order temporal regularization constrain the update. Translation and rotation corrections remain within 3 cm and 20° of the input trajectory.

Participant-specific hand calibration. For each hand, a flat-hand calibration sequence provides metric triangulated positions for a 21-joint skeleton. An initial MANO shape $\boldsymbol{\beta}_0$ (Romero et al., 2017) is calibrated against these observations. Let $\mathcal{B}$ contain the 20 skeleton bones, $\mathbf{X}_a^{\text{tri}}$ be triangulated joint a, and $\mathbf{J}_a(\boldsymbol{\beta}_0)$ the corresponding model joint. Physical scale is

$$s_{\text{hand}} = \frac{\sum_{(a,b)\in\mathcal{B}} \|\mathbf{X}_b^{\text{tri}} - \mathbf{X}_a^{\text{tri}}\|_2}{\sum_{(a,b)\in\mathcal{B}} \|\mathbf{J}_b(\boldsymbol{\beta}_0) - \mathbf{J}_a(\boldsymbol{\beta}_0)\|_2}. \tag{B.4}$$

The explicit scale is retained because MANO shape coefficients are not a global scale parameter. Components requiring a MANO-only shape vector receive an approximate scale-adjusted shape from ridge fitting in centered canonical vertex space. The participant-specific calibration is reused across sequences.

Hand-motion reconstruction. WiLoR (Potamias et al., 2025) provides 21 image-space keypoints per hand and view. For each joint, valid two-view pairs generate triangulation hypotheses. Reprojection residuals below 50 pixels define inliers, with at least three views required. The selected hypothesis is re-triangulated using all inliers. At the reconstruction stage's 10 fps sampling rate, each joint is filtered with a constant-velocity state $[x, y, z, v_x, v_y, v_z]^\top$, where x, y, z are coordinates and v_x, v_y, v_z their velocities. Measurement and acceleration-process standard deviations are 0.01 m and $0.12\,\text{m/s}^2$, respectively; missing observations use prediction only. This reconstruction sampling rate is distinct from the 30 fps capture rate. MANO fitting combines filtered metric joints, multi-view image keypoints, and calibrated projections, keeping hand shape fixed while estimating pose and global motion.

Joint hand–object refinement. The hand and object estimates are refined against shared metric depth. Variables are per-frame hand pose and global transforms, object pose, and one object-scale variable shared by the sequence. The objective is

$$\mathcal{L} = 10\mathcal{L}_{\text{SDF}} + \mathcal{L}_{\text{reg}}^h + \mathcal{L}_{\text{reg}}^o + \mathcal{L}_{\text{smooth}}^h + \mathcal{L}_{\text{smooth}}^o. \tag{B.5}$$

Here $\mathcal{L}_{\text{SDF}}$ aligns reconstructed geometry to a signed distance field (SDF) built from depth; $\mathcal{L}_{\text{reg}}^h$ and $\mathcal{L}_{\text{reg}}^o$ preserve hand and object initialization, and the two smoothness terms regularize their temporal evolution. The MeshSDF formulation follows HO-Cap (Wang et al., 2025a), using 32 depth views in the interaction region and 3 mm voxels. Optimization first uses priors and smoothness, then activates SDF alignment, and finally refines object scale with poses fixed. It does not include explicit contact, penetration, collision, or snapping losses.

Export representation. The standardized export uses a right-handed, $+Z$-up world frame with its recorded 4×4 transform from the reconstruction frame. The tabletop defines $z = 0$, estimated from the fifth percentile of the object bottom. MANO uses 45 axis–angle pose coordinates without PCA, with a flat-hand mean and no added mean-pose offset. Object vertices retain their canonical coordinates without recentering, axis flips, or mirroring.

B.3 Kinematic Retargeting and Replay Screening

Retargeting follows position-based kinematic fitting (Handa et al., 2020; Qin et al., 2023; Liu et al., 2025). For each frame and hand, fifteen target entries are fitted to corresponding Wuji-hand links: five MANO fingertips (indices $4, 8, 12, 16, 20$), five additional finger keypoints ($3, 7, 11, 15, 19$), and the wrist (index 0) repeated five times. The fit uses SmoothL1 position loss with transition scale 0.02 m and NLopt optimization with robot joint limits as hard bounds. The implementation uses a joint-limit tolerance of 10^{-3}, initializes from the preceding frame's solution, and uses inter-frame regularization parameter `norm_delta`$= 4 \times 10^{-3}$.

At conversion, source motion is interpreted at 10 fps and converted to 30 fps control references (`SOURCE_FPS=10`, `FPS_OUT=30`, `RESAMPLE_STRIDE=3`). The source visualization videos and per-clip provenance determine this timing; the export metadata's nominal 30 fps field is not used as the source rate. Pre-retargeting cleanup uses a 0.10 m per-frame displacement threshold at 30 fps, at most three despiking passes, and nearest-neighbor filling. These settings belong to conversion cleanup, not reconstruction filtering.

Trajectory cleanup occurs before retargeting. The retargeter's low-pass coefficient is 1, so it performs no additional low-pass smoothing. Wrist translation is computed as the MANO global translation plus its shape-dependent root joint, $\mathbf{t}+\mathbf{J}_0(\boldsymbol{\beta})$; a fixed frame correction aligns global orientation with the robot convention. The output pairs robot wrist/finger configurations with the reconstructed object trajectory.

Open-loop replay with zero policy correction classifies each candidate reference into six outcomes: immediate object loss (A), object ejection (B), contact without lifting (C), partial execution (D), tracking with an incorrect release or ending (E), and complete execution (F). Categories A and B are excluded, whereas C–F remain available for closed-loop training. Category C is retained because failure to lift under reference-only replay does not rule out recovery by feedback control. Screening determines reference eligibility; it is distinct from the verification used to admit executable demonstrations.

C Method and Training Details

C.1 Controller Interface

A reference clip m contains robot configurations $\hat{\mathbf{q}}_t^h \in \mathbb{R}^{26}$ for hand $h \in \{L, R\}$, object position $\hat{\mathbf{p}}_t^o \in \mathbb{R}^3$, and object rotation $\hat{R}_t^o \in SO(3)$ at control step t. Hats denote references; quantities without hats denote simulated states. Each hand has six wrist coordinates and twenty finger joints. The shared policy $\pi_\theta(\mathbf{a}_t \mid \mathbf{o}_t)$, with parameters θ, receives observation $\mathbf{o}_t$ and produces a 52-dimensional action $\mathbf{a}_t$ for both hands.

The 3,016-dimensional observation combines 2,884 base features with a 132-dimensional object descriptor: 128 basis-point-set features (Prokudin et al., 2019), scale, and centroid. Per-hand blocks contain proprioception, object pose and velocity, upcoming reference targets, state errors, and fingertip-to-object distance vectors in each wrist heading frame. Hand identity and left-hand reward weight are included. Forty contact-force channels are set to zero; contact forces contribute to reward supervision, not policy input.

Concatenating both reference configurations as $\hat{\mathbf{q}}_{t+1}$, the proportional–derivative (PD) target is

$$\mathbf{q}_t^{\mathrm{pd}} = \mathrm{clamp}(\hat{\mathbf{q}}_{t+1} + \mathbf{s} \odot \mathbf{a}_t, \mathbf{q}_{\min}, \mathbf{q}_{\max}), \qquad \mathbf{a}_t \in [-1, 1]^{52}. \tag{C.1}$$

The vector $\mathbf{s}$ sets coordinate-specific correction bounds, $\odot$ denotes element-wise multiplication, and $\mathbf{q}_{\min}, \mathbf{q}_{\max}$ are coordinate limits. Bounds are 0.1 m for wrist translation, 1.0 rad for wrist rotation, and 0.15 rad for finger joints. Corrections are applied around the next reference at every step and do not accumulate. This interface is inherited from reference-tracking approaches (Liu et al., 2025; Xu et al., 2025).

C.2 Reward Definition and Coefficients

We combine hand tracking, object motion, per-hand contact, and contact maintenance:

$$r_t = (1 + w_L)\big[\beta_{\mathrm{hand}} r_t^{\mathrm{hand}} + \beta_{\mathrm{obj}} r_t^{\mathrm{obj}} + \beta_{\mathrm{cont}} r_t^{\mathrm{cont}} + \beta_{\mathrm{hold}} r_t^{\mathrm{hold}}\big]. \tag{C.2}$$

The fixed hand weights are $w_L = 0.5$ and $w_R = 1$; adaptive hand weighting is disabled. The β coefficients sum to one. The inherited hand-tracking term uses Dexplore's per-hand multiplicative chain for wrist, finger-joint, and keypoint agreement (Xu et al., 2025), with the released subterm coefficients unchanged. Its two per-hand values are combined as $(r_t^{\mathrm{hand},R} + w_L r_t^{\mathrm{hand},L})/(1 + w_L)$. Object-position and object-rotation factors within this hand chain have zero weights (`reward.op=reward.or=0`), avoiding double counting with the separate object reward.

For object linear velocity $\mathbf{v}_t^o$ and reference velocity $\hat{\mathbf{v}}_t^o$ from finite differences, the object reward is

$$r_t^{\mathrm{obj}} = e^{-\kappa_p \|\mathbf{p}_t^o - \hat{\mathbf{p}}_t^o\|_2} e^{-\kappa_R d_{\mathrm{SO}(3)}(R_t^o, \hat{R}_t^o)} e^{-\kappa_v \|\mathbf{v}_t^o - \hat{\mathbf{v}}_t^o\|_2}. \tag{C.3}$$

The implemented object reward multiplies the three factors and has no η mixture weights. Joint position, orientation, and velocity agreement is rewarded; the velocity factor penalizes holding a

Table S2: Reward and generalist-training settings. Distances in reward exponents use metres, rotations radians, and velocities metres per second.

Parameter	Setting
$(\beta_{\text{hand}}, \beta_{\text{obj}}, \beta_{\text{cont}}, \beta_{\text{hold}})$	$(0.25, 0.45, 0.15, 0.15)$
$(\kappa_p, \kappa_R, \kappa_v)$	$(9, 2, 14)$
$(f_{\min}, \delta_{\max})$	$(0.5\,\text{N}, 0.03\,\text{m})$
(κ_c, ζ)	$(30\,\text{m}^{-1}, 0.5)$
Parallel simulation environments	8,192
BC coefficient / pairs per minibatch	50 / 4,096
Recovery / perturbation curriculum	Disabled
Fixed hand weights (w_L, w_R)	$(0.5, 1)$

moving reference stationary. This term is independent of contact gating. The positive coefficients $(\kappa_p, \kappa_R, \kappa_v) = (9, 2, 14)$ control the three error sensitivities. The full reward still adds the hand, object, contact, and hold terms.

The reference-required set $\mathcal{C}_t^h$ uses the five fingertip slots $[3, 6, 9, 12, 15]$ within reference contact columns [206:222] for the left hand and [222:238] for the right. A fingertip is required when its channel exceeds 0.01; hand activity uses an element-wise logical OR over these tests, not a sum of signed labels. This reference mask is distinct from simulated contact. For fingertip i, surface distance $\delta_{i,t}^h$ and simulated net force magnitude $f_{i,t}^h$ define a contact proxy

$$\alpha_{i,t}^h = \mathbf{1}[f_{i,t}^h > f_{\min}]\mathbf{1}[\delta_{i,t}^h < \delta_{\max}]. \tag{C.4}$$

For nonempty $\mathcal{C}_t^h$, proximity and instantaneous required-contact coverage give

$$\begin{aligned} r_t^{\text{cont},h} &= \frac{1}{|\mathcal{C}_t^h|} \sum_{i \in \mathcal{C}_t^h} \left[\zeta e^{-\kappa_c \delta_{i,t}^h} + (1-\zeta)\alpha_{i,t}^h \right], \\ r_t^{\text{hold},h} &= \frac{1}{|\mathcal{C}_t^h|} \sum_{i \in \mathcal{C}_t^h} \alpha_{i,t}^h. \end{aligned} \tag{C.5}$$

The set of engaged hands is $\mathcal{E}_t = \{h : \mathcal{C}_t^h \neq \varnothing\}$. Each contact-related term $u \in \{\text{cont}, \text{hold}\}$ is aggregated as

$$r_t^u = \begin{cases} \dfrac{\sum_{h \in \mathcal{E}_t} w_h r_t^{u,h}}{\sum_{h \in \mathcal{E}_t} w_h}, & \mathcal{E}_t \neq \varnothing, \\ 0, & \mathcal{E}_t = \varnothing. \end{cases} \tag{C.6}$$

Thus an inactive hand contributes no contact reward, and an unsuccessful engaged hand does not multiply away the other hand's signal. The force–distance rule remains a proxy because net force does not identify the collision partner. Contact maintenance measures instantaneous coverage, not a no-slip condition, allowing valid sliding and rolling. Proximity guides fingertips before contact; the hold term evaluates required-contact coverage throughout the reference contact interval.

C.3 Clip-Centered Advantages and PPO+BC

We use proximal policy optimization (PPO) (Schulman et al., 2017) with generalized advantage estimation (GAE) (Schulman et al., 2016), and add behavioral cloning (BC) from verified demonstrations. Let $V_\phi(\mathbf{o}_t)$ denote the learned value estimate, γ the discount factor, and λ_{GAE} the temporal trace parameter. Writing b_t for the bootstrap-validity mask, the temporal-difference residual is

$$\Delta_t = r_t + \gamma b_t V_\phi(\mathbf{o}_{t+1}) - V_\phi(\mathbf{o}_t). \tag{C.7}$$

GAE accumulates these residuals with weights $(\gamma\lambda_{\text{GAE}})^\ell$, stopping the trace at episode boundaries. Bootstrap and timeout handling follow the base implementation. This temporal estimate is computed before any clip-level adjustment.

For an on-policy batch $\mathcal{B}$, transition index j, and reference identity $m(j)$, define $\mathcal{G}_m = \{j \in \mathcal{B} : m(j) = m\}$. CCA uses

$$\bar{A}_m = \frac{1}{|\mathcal{G}_m|} \sum_{j \in \mathcal{G}_m} A_j^{\text{GAE}}, \qquad \widetilde{A}_j = A_j^{\text{GAE}} - \bar{A}_{m(j)}. \tag{C.8}$$

Each nonempty group has zero empirical mean after centering. For two transitions j, k from the same clip, $\widetilde{A}_j - \widetilde{A}_k = A_j^{\mathrm{GAE}} - A_k^{\mathrm{GAE}}$. These identities explain the intended behavior: remove a common reference-dependent offset while preserving within-reference differences. CCA is a mean shift, not division by a clip standard deviation, and it retains the learned critic and temporal GAE calculation. Its empirical centering changes the actor update; the algebraic identities alone do not establish an unbiased policy-gradient estimator.

Let θ_{old} denote the policy that collected $\mathcal{B}$, ϵ the PPO clipping radius, and $\rho_j(\theta) = \pi_\theta(\mathbf{a}_j \mid \mathbf{o}_j)/\pi_{\theta_{\mathrm{old}}}(\mathbf{a}_j \mid \mathbf{o}_j)$. The actor loss is

$$\mathcal{L}_{\mathrm{PPO}} = -\mathbb{E}_{j\sim\mathcal{B}}\big[\min\big(\rho_j(\theta)\widetilde{A}_j, \mathrm{clip}(\rho_j(\theta), 1-\epsilon, 1+\epsilon)\widetilde{A}_j\big)\big]. \tag{C.9}$$

For demonstration library $\mathcal{D}$ and policy mean μ_θ, the combined objective is

$$\mathcal{L}_{\mathrm{actor}} = \mathcal{L}_{\mathrm{PPO}} + \lambda_{\mathrm{BC}}\mathbb{E}_{(\mathbf{o}^D,\mathbf{a}^D)\sim\mathcal{D}}\|\mu_\theta(\mathbf{o}^D) - \mathbf{a}^D\|_2^2, \qquad \lambda_{\mathrm{BC}} = 50. \tag{C.10}$$

The superscript D denotes a stored observation–action pair. Demonstrations use a separate sampling budget and do not enter the PPO probability ratio or GAE. The critic retains its value-regression objective.

C.4 Demonstration Collection, Sampling, and Compatibility

Following DexTrack's reinforcement–imitation loop (Liu et al., 2025), specialist training covers all references retained by replay screening. Each trajectory-specific controller seeks an executable realization of its reference; successful rollouts are exported and replay-verified before admission to the demonstration library. The shared controller is periodically evaluated on the training pool, and its verified successes supplement the library for subsequent PPO+BC updates. Held-out references are excluded throughout.

Verification checks object displacement and, when required by the reference, lifting. An exported rollout must succeed again on replay. This admission procedure selects executable supervision and is distinct from the trajectory-fidelity metrics in Section D.2. Each BC minibatch samples 4,096 observation–action pairs independently of the PPO minibatch.

A versioned action/reference contract records the action parameterization, reference version, and retargeting fingerprint. Changes to this contract invalidate incompatible stored actions.

C.5 Optimization and Simulation Settings

The generalist recipe uses Isaac Gym (Makoviychuk et al., 2021) with 8,192 environments. Actor and critic use separate multilayer perceptrons with widths $[1024, 1024, 1024, 512]$ and ReLU activations. The Gaussian policy uses a fixed, non-learned scale parameter configured by `sigma_init=-2.9`. Input, value, and advantage normalization and value bootstrapping are enabled.

Demonstration supervision uses $\lambda_{\mathrm{BC}} = 50$ with a separate minibatch budget. Adaptive recovery sampling and object-position perturbation curricula are disabled. Checkpoints are saved every 25 training iterations.

D Experimental Protocols and Additional Details

D.1 Evaluation Splits

The public-only model uses $1{,}845$ references and Ours+BiHOI uses $3{,}420$, including all $1{,}575$ BiHOI sequences. Evaluation comprises 53 new trajectories of training objects (UNSEEN-TRAJ) and 164 clips involving held-out objects (UNSEEN-OBJ). Neither split contributes to policy optimization, specialist training, or demonstration collection. Each method uses one checkpoint across both splits without fine-tuning.

The main DONIC results use the frozen p2t checkpoint. Evaluation uses the policy mean and starts clips from their first frame. Neither evaluation split supplies training transitions or demonstrations.

Table S3: Optimization and simulation configuration.

Setting	Value
Optimizer / learning rate / schedule	Adam / 2×10^{-5} / constant
Rollout horizon / PPO minibatch	64 / 16,384
PPO passes per rollout / clipping radius	6 / 0.2
Discount / GAE trace parameter	0.99 / 0.95
Value / entropy / action-bound loss coefficients	5 / 0 / 10
Value clipping / gradient truncation	Disabled / disabled
Configured gradient norm / seed	1.0 / 42
Maximum episode steps	2,000
Control-frequency inverse / simulation substeps	2 / 4
Object density / hand friction	400 / 2.0
Linear / angular object-velocity caps	20 / 50
Reset angular-velocity cap	20
`PREGRASP_INIT`	0.9
Reference feed-forward / deployment mask	Enabled / enabled
Reference-correction bounds (wrist trans./rot.; fingers)	0.1 m / 1.0 rad / 0.15 rad
BC coefficient / demonstration batch	50 / 4,096
Recovery / object-perturbation curriculum	Disabled
Environment transitions per rollout iteration	$8{,}192 \times 64 = 524{,}288$

D.2 Bimanual Tracking Metrics

We extend AdaDexTrack's trajectory-fidelity evaluation (Adalibieke et al., 2026) to require both hands to satisfy the tracking criteria. Let $\mathcal{M}$ be the evaluated clips, T_m the number of evaluated frames in clip m, and $h \in \{L, R\}$ a hand. The following definitions apply to each frame; clip indices are omitted when unambiguous.

Object, wrist, and finger errors. Object translation error $e^o_{p,t}$ is the Euclidean distance between simulated and reference positions. For unit object-orientation quaternions $\mathbf{u}^o_t$ and $\hat{\mathbf{u}}^o_t$, rotation error is

$$e^o_{R,t} = 2 \arccos\left(|\langle \mathbf{u}^o_t, \hat{\mathbf{u}}^o_t \rangle|\right). \tag{D.1}$$

The absolute inner product accounts for the quaternion double cover, giving an angle in $[0, \pi]$.

For each hand, the same robot Unified Robot Description Format (URDF) forward kinematics maps simulated and retargeted reference configurations to wrist poses $(\mathbf{p}^h_t, R^h_t)$ and $(\hat{\mathbf{p}}^h_t, \hat{R}^h_t)$. The wrist errors are

$$e^h_{p,t} = \|\mathbf{p}^h_t - \hat{\mathbf{p}}^h_t\|_2, \qquad e^h_{R,t} = \arccos\left(\frac{\mathrm{tr}((R^h_t)^\top \hat{R}^h_t) - 1}{2}\right). \tag{D.2}$$

The source human wrist is not used in this comparison: the metric measures tracking after retargeting, rather than adding human-to-robot retargeting mismatch. With finger-joint angles $q^h_{t,k}$ and $\hat{q}^h_{t,k}$, finger error is

$$e^h_{f,t} = \frac{1}{20} \sum_{k=1}^{20} |q^h_{t,k} - \hat{q}^h_{t,k}|. \tag{D.3}$$

The six virtual wrist coordinates are excluded. All angular errors and thresholds are computed in radians; translation errors are expressed in centimetres for reporting and thresholding.

Aggregation of reported errors. For any frame error e, write $\bar{e}_m = T_m^{-1} \sum_t e_{m,t}$ for its trajectory mean. The reported object metrics are

$$T_o = \frac{1}{|\mathcal{M}|} \sum_m \bar{e}^o_{p,m}, \qquad R_o = \frac{1}{|\mathcal{M}|} \sum_m \bar{e}^o_{R,m}. \tag{D.4}$$

Hand metrics first take the worse of the two per-hand trajectory means, then average over clips:

$$(T_h, R_h, E_{\mathrm{finger}}) = \frac{1}{|\mathcal{M}|} \sum_m \left(\max_h \bar{e}^h_{p,m}, \max_h \bar{e}^h_{R,m}, \max_h \bar{e}^h_{f,m} \right). \tag{D.5}$$

The worse hand can differ by metric. In particular, this is not an average across hands or a framewise maximum before temporal averaging.

Success rate and completion. For object errors x_p^o, x_R^o and hand errors x_p^h, x_R^h, x_f^h, define

$$\mathcal{K}(x^o, x^h) = \mathbf{1}\left[\frac{x_p^o + x_p^h}{2} \le 10\,\text{cm}\right] \mathbf{1}\left[\frac{x_R^o + x_R^h}{2} \le \frac{\pi}{3}\right] \mathbf{1}\left[x_f^h \le \frac{2\pi}{9}\right]. \tag{D.6}$$

The angular thresholds $\pi/3$ and $2\pi/9$ correspond to 60° and 40°. Let $\bar{e}_m^o$ and $\bar{e}_m^h$ collect the relevant trajectory means. Mean-error success rate is

$$\text{SR} = \frac{100}{|\mathcal{M}|} \sum_m \prod_{h \in \{L,R\}} \mathcal{K}(\bar{e}_m^o, \bar{e}_m^h). \tag{D.7}$$

The product implements a logical AND: both hands must independently pass using their own trajectory means. Completion instead applies the same tests separately at every frame:

$$\text{Comp} = \frac{100}{|\mathcal{M}|} \sum_m \frac{1}{T_m} \sum_t \prod_{h \in \{L,R\}} \mathcal{K}(e_{m,t}^o, e_{m,t}^h). \tag{D.8}$$

A later valid frame contributes even after an earlier violation. Comp therefore measures per-frame satisfaction, not the length of the prefix before the first failure. Both metrics weight trajectories equally rather than weighting longer clips more heavily.

D.3 Baseline Implementations and Comparison Scope

DexTrack-bi. We use DexTrack's released specialist–distillation pipeline (Liu et al., 2025), adapting its observation and action interfaces to two Wuji hands. Trajectory-specific controllers produce demonstrations for generalist distillation, and evaluation follows the released per-trajectory protocol.

Dexplore-bi. We run the released Dexplore code (Xu et al., 2025) with two-hand observation and action interfaces, retaining its original reward and reference-scoped exploration.

Evaluation settings. All methods use Isaac Gym, two Wuji hands, the same public training clips, and the same held-out split identities. Tracking errors are computed using the bimanual metrics defined in Section D.2. All 164 unseen-object and 53 unseen-trajectory clips are included.

D.4 Ablation Definitions and Additional Data

The learning-signal ablations in the main table examine CCA and object-centered supervision. The w/o CCA variant removes the clip-wise mean subtraction after GAE. The w/o contact gate variant disables `DEXPLORE_CG_REWARD`, replacing the simulated force-and-distance contact proxy by the force-only test $\mathbf{1}[f_{i,t}^h > 0.5]$. It removes the 3 cm contact-ownership condition, not the object-motion reward. Reward coefficients β and fixed hand weights are unchanged; they are not renormalized. Ours+BiHOI tests the effect of adding all $1{,}575$ captured sequences to the public pool while retaining the same public evaluation splits.

Ablation outcomes. Removing either learning signal lowers both SR and Comp on both splits. CCA removal also increases all five tracking errors; removing the contact gate improves some individual errors but reduces joint bimanual success. Adding BiHOI improves all seven metrics on both splits, reaching SR of 88.68% and 59.76% on unseen trajectories and objects, respectively. Unseen-object translation error falls from 15.74 to 13.96 cm.

The learning-signal ablations retain the full models training data and optimization settings, changing only the indicated component. Ours+BiHOI uses the same training recipe with the expanded reference pool.

D.5 Qualitative Comparisons

Figure S2 compares five bimanual manipulation sequences. Across these examples, DONIC more closely reproduces reference hand–object coordination and object motion, whereas the baselines exhibit limited object motion or loss of hand–object alignment.

D.6 Limitations

The quantitative experiments evaluate reference-conditioned, single-rigid-object tracking in simulation. Unconstrained goal-conditioned manipulation and held-out operation-category generalization are outside this evaluation. The capture inventory includes broader object types than the rigid-object representation described in Section B.1. Contact supervision uses a net-force/proximity proxy, and reconstruction has no explicit contact or collision loss.

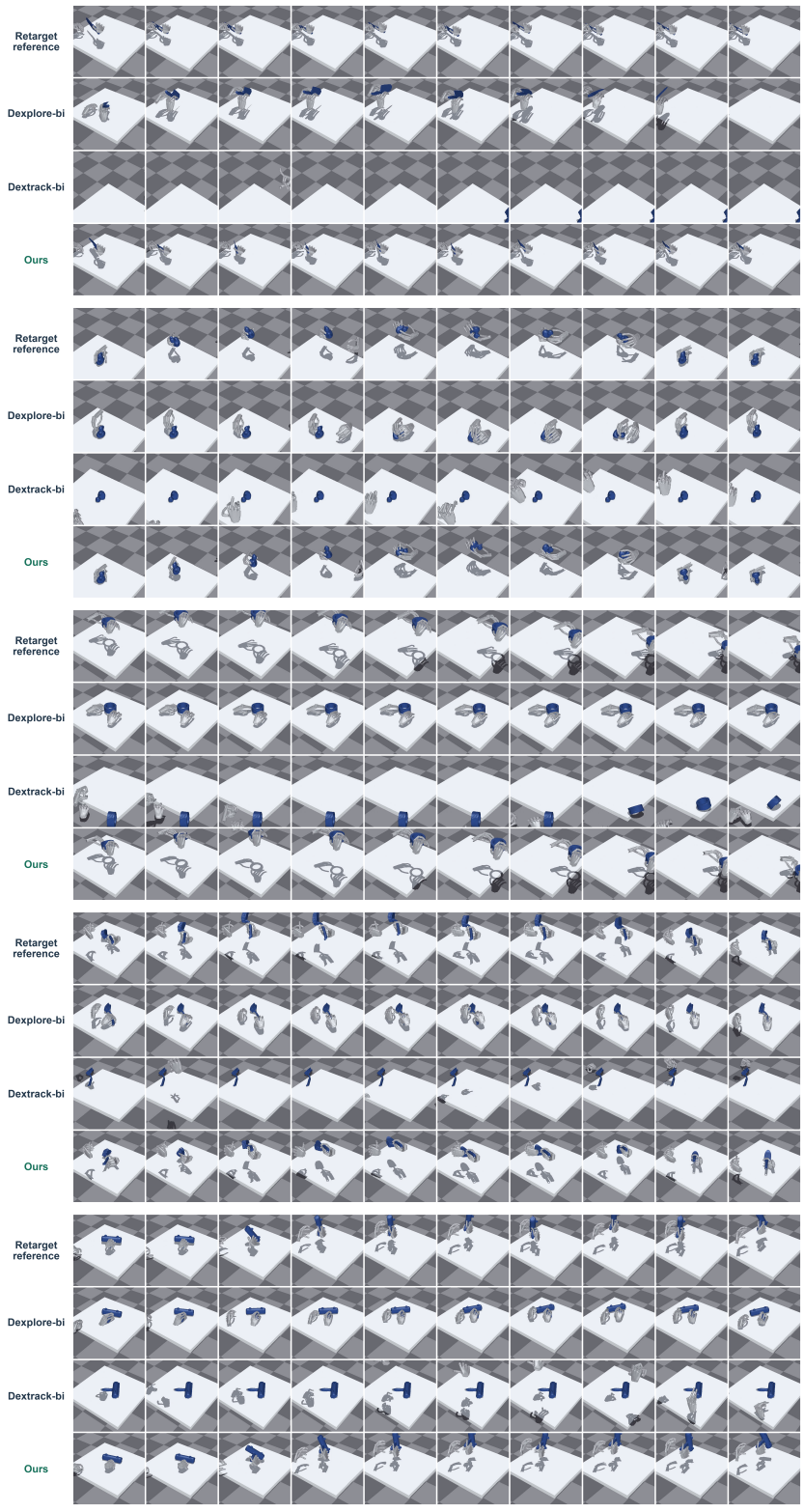

Figure S2: **Qualitative bimanual tracking comparisons.** Rows in each group show the reference, Dexplore-bi, DexTrack-bi, and DONIC (Ours); time runs left to right.