

[illegible]

A. POLICY AND ENVIRONMENT DETAILS

$$\mathbf{e}_{i,t,k}^{b,x} = \hat{\mathbf{x}}_{i,k}^b - \mathbf{x}_t^b. \quad (\text{S1})$$
$$e_{i,t,k}^{b,Q} = Q_t^b \otimes (\hat{Q}_{i,k}^b)^{-1}. \quad (\text{S2})$$

A.1. Observation Space

$$\mathcal{H} = \{0, -1, -2, -4, -8\}, \quad \mathcal{F} = \{0, 1, 2, 4, 8\}. \quad (\text{S3})$$
$$h_t = [q_t, \cos q_t, \sin q_t, p_t^w, Q_t^w, v_t^w, \omega_t^w]. \quad (\text{S4})$$
$$\mathbf{c}_t = [\dot{\mathbf{q}}_t, \mathbf{c}_t^{\text{obj}}, \mathbf{c}_t^{\text{tip}}]. \quad (\text{S5})$$

$$\mathbf{c}_t^{\text{obj}} = [\mathbf{p}_t^o - \mathbf{p}_t^w, \mathbf{p}_t^o, \mathbf{Q}_t^o, \mathbf{v}_t^o, \boldsymbol{\omega}_t^o, \mathbf{p}_t^{\text{CoM}} - \mathbf{p}_t^w, w_o]. \quad (\text{S6})$$

$$\mathbf{c}_t^{\text{tip}} = \{\mathbf{f}_t^j, \|\mathbf{f}_t^j\|_2\}_{j=1}^5. \quad (\text{S7})$$

$$\mathbf{r}_{i,t,k}^b = [\mathbf{r}_{i,t,k}^{b,\text{lin}}, \mathbf{r}_{i,t,k}^{b,\text{rot}}]. \quad (\text{S8})$$

$$\mathbf{r}_{i,t,k}^{b,\text{lin}} = [\mathbf{e}_{i,t,k}^{b,p}, \hat{\mathbf{v}}_{i,k}^b, \mathbf{e}_{i,t,k}^{b,v}]. \quad (\text{S9})$$

$$\mathbf{r}_{i,t,k}^{b,\text{rot}} = [\hat{Q}_{i,k}^b, e_{i,t,k}^{b,Q}, \hat{\omega}_{i,k}^b, e_{i,t,k}^{b,\omega}]. \quad (\text{S10})$$

$$\mathbf{r}_{i,t,k}^{\text{link}} = \{\hat{\mathbf{p}}_{i,k}^j - \mathbf{p}_t^j, \hat{\mathbf{v}}_{i,k}^j, \hat{\mathbf{v}}_{i,k}^j - \mathbf{v}_t^j\}_{j=1}^J. \quad (\text{S11})$$
$$\mathbf{r}_{i,t,k} = [\mathbf{r}_{i,t,k}^w, \mathbf{r}_{i,t,k}^o, \mathbf{r}_{i,t,k}^{\text{link}}, \hat{\mathbf{d}}_{i,k}^{\text{tip}}]. \quad (\text{S12})$$
$$o_t = [h_t^{\mathcal{H}}, c_t^{\mathcal{H}}, r_{i,t}^{\mathcal{F}}, d_t^o, b_i^o, a_t^{\text{base}}, \ell_t]. \quad (\text{S13})$$
$$\mathbf{c}_t = [\dot{\mathbf{q}}_t, \mathbf{c}_t^{\text{obj}}, \mathbf{c}_t^{\text{tip}}]. \quad (\text{S5})$$

A.2. Action Space

For n actuated finger joints, the action comprises normalized wrist translation, wrist rotation, and finger target residuals:

$$\mathbf{a}_t = [\mathbf{a}_t^p, \mathbf{a}_t^R, \mathbf{a}_t^q] \in [-1, 1]^{6+n}, \quad \boldsymbol{\delta}_t = 2\mathbf{a}_t. \quad (\text{S14})$$

The kinematic base action comes from the retargeted hand reference. Its wrist position error $\mathbf{e}_{i,t,t}^{w,p}$ is clipped to 0.10 m; the axis-angle error of $\hat{\mathbf{Q}}_{i,t}^w \otimes (\mathbf{Q}_t^w)^{-1}$ is clipped to 1 rad. The observation uses the reverse quaternion difference in Eq. (S2). The finger base $\mathbf{b}_t^q$ is the normalized retargeted target.

In wrist PID control, residuals augment the base force and torque:

$$\tilde{\mathbf{F}}_t = \text{PID}_p(\bar{\mathbf{e}}_t^p) + 500\Delta t s_p \boldsymbol{\delta}_t^p. \quad (\text{S15})$$

$$\tilde{\mathbf{M}}_t = \text{PID}_R(\bar{\mathbf{e}}_t^R) + 200\Delta t s_R \boldsymbol{\delta}_t^R. \quad (\text{S16})$$

Here $\bar{\mathbf{e}}_t^p$ and $\bar{\mathbf{e}}_t^R$ denote the clipped errors, $\Delta t = 1/60$ s, $s_p = 1$, and $s_R = 0.1$; PID gains depend on the hand model. Finger residuals are added in normalized joint coordinates and mapped to joint limits:

$$\mathbf{u}_t^q = \text{clip}(\mathbf{b}_t^q + \boldsymbol{\delta}_t^q, -1, 1). \quad (\text{S17})$$

$$\mathbf{q}_t^{\text{raw}} = \mathbf{q}_{\min} + \frac{1}{2}(\mathbf{u}_t^q + 1) \odot (\mathbf{q}_{\max} - \mathbf{q}_{\min}). \quad (\text{S18})$$

Wrist force, torque, and finger targets use temporal smoothing with $\alpha = 0.4$:

$$\mathbf{y}_t = \alpha \tilde{\mathbf{y}}_t + (1 - \alpha) \mathbf{y}_{t-1}, \quad \mathbf{y} \in \{\mathbf{F}, \mathbf{M}, \mathbf{q}^{\text{cmd}}\}. \quad (\text{S19})$$

For finger control, $\tilde{\mathbf{q}}_t^{\text{cmd}} = \mathbf{q}_t^{\text{raw}}$; the smoothed target is clipped to joint limits. Wrist force $\mathbf{F}_t$ (N) and torque $\mathbf{M}_t$ (N·m) act on the wrist rigid body; finger targets are in radians. Object motion follows contact dynamics. Bimanual actions concatenate both hands' residuals.

A.3. Reward Function

The task-agnostic reward comprises hand pose, object pose, velocity, reference-guided contact, and mechanical-power terms. For the set $\mathcal{G}$ of these groups,

$$r_t = \sum_{g \in \mathcal{G}} \lambda_g r_t^g. \quad (\text{S20})$$

The hand group includes wrist and finger tracking. Table S1 specifies each component's weight; $\lambda_g = 1$ for all five groups.

At the current reference frame, pose errors are

$$E_{b,p} = \|\mathbf{e}_{i,t,t}^{b,p}\|_2, \quad E_{b,R} = 2 \arccos |\langle \mathbf{Q}_t^b, \hat{\mathbf{Q}}_{i,t}^b \rangle|. \quad (\text{S21})$$

For each hand-link group G ,

$$E_G = |\mathcal{I}_G|^{-1} \sum_{j \in \mathcal{I}_G} \|\hat{\mathbf{p}}_{i,t}^j - \mathbf{p}_t^j\|_2. \quad (\text{S22})$$

Velocity errors average absolute component differences:

$$E_{b,v} = \frac{1}{3} \|\mathbf{e}_{i,t,t}^{b,v}\|_1, \quad E_{b,\omega} = \frac{1}{3} \|\mathbf{e}_{i,t,t}^{b,\omega}\|_1. \quad (\text{S23})$$

Term (unit)	Component reward	Weight
Hand pose tracking		
Wrist position (m)	$\exp(-40E_{w,p})$	0.10
Wrist orientation (rad)	$\exp(-E_{w,R})$	0.60
Thumb (m)	$\exp(-100E_T)$	0.90
Index (m)	$\exp(-90E_I)$	0.80
Middle (m)	$\exp(-80E_M)$	0.75
Ring (m)	$\exp(-60E_R)$	0.60
Little (m)	$\exp(-60E_P)$	0.60
Link group L_1 (m)	$\exp(-50E_{L_1})$	0.50
Link group L_2 (m)	$\exp(-40E_{L_2})$	0.30
Object pose tracking		
Object position (m)	$\exp(-80E_{o,p})$	5.00
Object orientation (rad)	$\exp(-6E_{o,R})$	1.00
Velocity matching		
Wrist linear (m/s)	$\exp(-E_{w,v})$	0.10
Wrist angular (rad/s)	$\exp(-E_{w,\omega})$	0.05
Hand links (m/s)	$\exp(-E_{\text{link},v})$	0.10
Object linear (m/s)	$\exp(-E_{o,v})$	0.10
Object angular (rad/s)	$\exp(-E_{o,\omega})$	0.10
Reference-guided contact		
Fingertip contact (N)	$\exp[-(F_t^{\text{tip}}/N + 10^{-5})^{-1}]$	1.00
Mechanical-power regularization		
Joint power (W)	$\exp(-10P_t^q)$	0.50
Wrist power (W)	$\exp(-2P_t^w)$	0.50

TABLE S1: Single-hand reward components. Units refer to the underlying errors or physical quantities.

Parameter	Value
PPO rollout / mini-epochs / minibatch	32 / 5 / 1024
Discount γ / GAE λ_{GAE}	0.99 / 0.95
Adam learning rate / PPO clip	5×10^{-4} / 0.2
Reset slots K / anchor probability α	3 / 0.1
Survival threshold $c_{\min}$ / score decay δ	0.1 / 10^{-5}
Bin width / lookback n / factor λ	15 / 5 / 0.8

TABLE S2: Training and reset allocation hyperparameters.

For tracked hand links,

$$E_{\text{link},v} = (3J)^{-1} \sum_{j=1}^J \|\hat{\mathbf{v}}_{i,t}^j - \mathbf{v}_t^j\|_1. \quad (\text{S24})$$

Reference fingertip-object distance gates the simulated fingertip force. For fingertip j , set $u_t^j = \text{clip}((0.03 - \hat{d}_{i,t}^{\text{tip},j})/0.01, 0, 1)$, with distances in metres, and let

$$F_t^{\text{tip}} = \sum_{j=1}^5 u_t^j \|\mathbf{f}_t^j\|_2. \quad (\text{S25})$$

Mechanical-power terms use

$$P_t^q = \sum_{j=1}^n |\tau_{t,j}^q \dot{q}_{t,j}|. \quad (\text{S26})$$

$$P_t^w = |\mathbf{F}_t^\top \mathbf{v}_t^w| + |\mathbf{M}_t^\top \boldsymbol{\omega}_t^w|. \quad (\text{S27})$$

Component rewards and weights are dimensionless; exponential scales assume SI units. Bimanual rewards sum per-hand terms, including an object term for each hand.

B. TRAINING AND RESET ALLOCATION DETAILS

B.1. PPO and Network Configuration

Each observation block is normalized using running statistics before concatenation. A feature MLP with three 512-unit Swish layers and a 256-dimensional projection feeds a

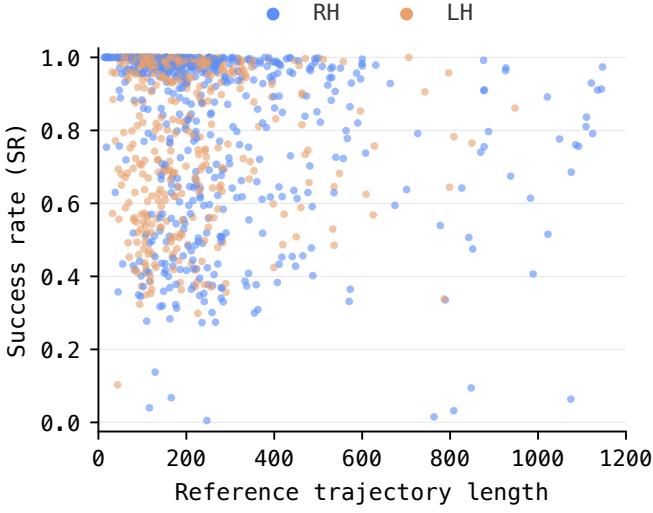


Fig. S2: Per-trajectory success rates versus reference trajectory length. Colors distinguish right-hand (RH) and left-hand (LH) trajectories.

shared ELU trunk of widths 256, 512, 128, and 64. Separate heads output the mean of a diagonal Gaussian policy and the scalar critic $V_\phi(o_t)$; the Gaussian has trainable, observation-independent log standard deviations. Policy samples are bounded to $[-1, 1]$ before residual action synthesis. Clipped PPO jointly updates the shared features and both heads using generalized advantage estimation. Value targets are normalized during optimization, while reset allocation uses critic estimates on the original reward scale.

B.2. Reset Buffer Update

For each reference phase (i, t) , the reset buffer $\mathcal{B}_{i,t}$ holds K scored states, initialized from the phase-aligned reference state $s_{i,t}^{\text{ref}}$ and its critic score. The fixed reference anchor is sampled through a separate α branch. Each state stores hand and object root configurations and velocities, finger joint positions and velocities, and controller targets. Before action execution at relative step u of episode e , the physical state is paired with its critic estimate. At episode end, the pair is assigned to the corresponding reference frame:

$$t = t_e^{\text{start}} + u, \quad v_{e,u} = V_\phi(o_{e,u}). \quad (\text{S28})$$

For candidates assigned to the same phase, scores are processed in descending order. Here $c(s)$ is the fraction of

remaining reference steps tracked from the candidate before episode termination. A candidate must satisfy $c(s) > c_{\min}$ and strictly exceed the current lowest buffer score to replace that entry; the lowest score is recomputed after each replacement. After a nonempty episode-finalization batch, candidate scores decay once as $v \leftarrow (1 - \delta)v$. The reference anchor remains fixed and is used if all candidate sampling weights vanish.

B.3. Reset Phase Allocation

Each reference trajectory is partitioned into temporal bins of 15 control frames; its final bin may be shorter. The bins are indexed globally across trajectories. For each rollout step with $\ell_t > 0$, the critic estimate on the original reward scale is normalized by the remaining discounted horizon:

$$\nu_t = \frac{V_\phi(o_t)}{\sum_{h=0}^{\ell_t-1} \gamma^h}. \quad (\text{S29})$$

For bins receiving samples, let $\mathcal{D}_b$ be the multiset of rollout steps assigned to bin b , including repeated visits. Their mean is $\bar{\nu}_b = |\mathcal{D}_b|^{-1} \sum_{t \in \mathcal{D}_b} \nu_t$. The difficulty $f_b = \nu_{\max} - \bar{\nu}_b$ uses the theoretical maximum normalized value implied by the reward. It propagates to preceding bins within the same trajectory:

$$\tilde{f}_b = \sum_{d=0}^{n-1} \lambda^d f_{b+d}, \quad (\text{S30})$$

with terms crossing the trajectory boundary omitted. The propagated score updates the phase priority by exponential moving average:

$$\bar{f}_b \leftarrow (1 - \beta)\bar{f}_b + \beta\tilde{f}_b. \quad (\text{S31})$$

Sampling normalizes $\bar{f}_b$ over all bins, then selects uniformly among the n_b valid frames of the chosen bin. Thus, for a frame (i, t) in bin $b(i, t)$,

$$p_{\text{phase}}(i, t) = \frac{\bar{f}_{b(i,t)}}{n_{b(i,t)} \sum_{j=1}^B \bar{f}_j}. \quad (\text{S32})$$

B.4. Hyperparameters

Table S2 summarizes the principal training and reset parameters.

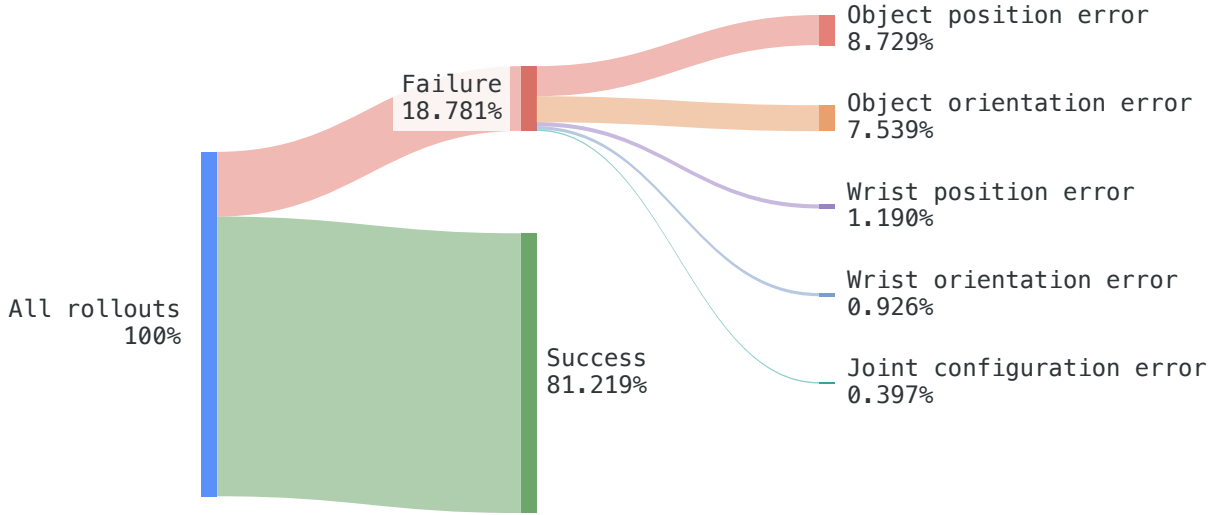


Fig. S3: Evaluation outcomes and estimated failure-category proportions. Failure labels denote tracking errors exceeding their respective thresholds. All percentages are relative to the 1,017 evaluation rollouts. Failure-category proportions are estimated by proportionally allocating unclassified failures across the five categories.

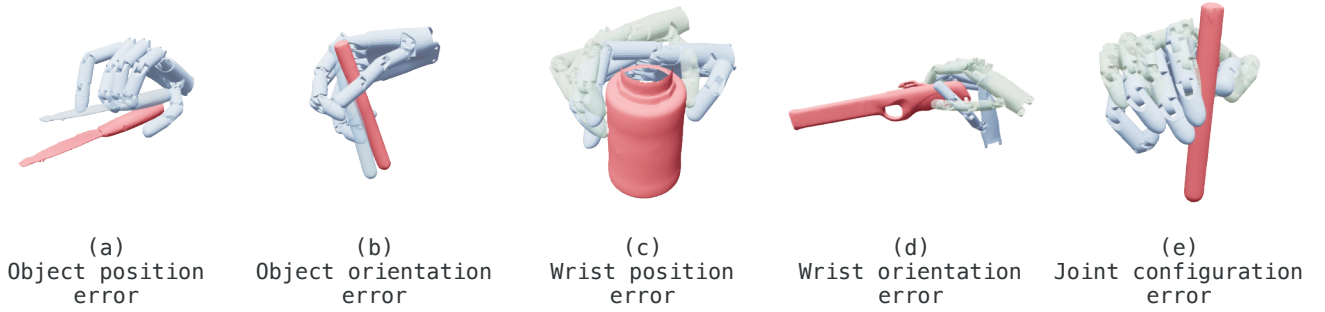


Fig. S4: Representative examples of the five failure categories. Translucent hands and objects indicate the reference.

C. ADDITIONAL RESULTS

C.1. Per-Trajectory Success Rates

We evaluate DexMimic on 1,302 trajectories from OakInk and HOT3D. Figure S2 plots success rate against reference trajectory length, with colors distinguishing right-hand (RH) and left-hand (LH) trajectories. Figure S1 shows representative states from 28 recorded DexMimic trajectories, including 23 single-hand and five bimanual examples.

C.2. Failure Case Analysis

Across 1,017 evaluation rollouts on a subset of OakInk trajectories, DexMimic succeeds in 826 cases (81.219%). Rollouts terminate early when tracking errors in object position, object orientation, wrist position, wrist orientation, or finger joint configuration exceed their respective thresholds. The estimated breakdown in Fig. S3 is dominated by object tracking failures, indicating that maintaining the reference object position and orientation remains the primary challenge in these evaluations. Figure S4 shows representative cases for the five categories.