Supplementary Appendix

A Reward Design

All reward terms with initial weights are detailed in Table 1. The overall reward design is divided into two categories: task-related rewards and regularization rewards.

Table 1: Reward terms and weights.

Term	Expression	Initial Weight θ	σ
	Task		
Joint position	$\exp(-\|\boldsymbol{q}_t - \boldsymbol{q}_{ref}\|_2^2/\sigma_{\text{jpos}})$	0.75	1.0
Joint velocity	$\exp(-\|\dot{\boldsymbol{q}}_t - \dot{\boldsymbol{q}}_{ref}\|_2^2/\sigma_{\text{jvel}})$	0.5	1.0
Body link position	$\exp(-\|\boldsymbol{q}_t^{link} - \boldsymbol{q}_{ref}^{link}\|_2^2/\sigma_{\text{pos}})$	1.0	0.03 (upper)/0.1 (lower)
Body link velocity	$\exp(-\|\dot{\boldsymbol{q}}_t^{link} - \dot{\boldsymbol{q}}_{ref}^{link}\|_2^2/\sigma_{\text{vel}})$	0.5	1.0
Body link rotation	$\exp(-\|\boldsymbol{\vartheta}^{link} \ominus \boldsymbol{\vartheta}_{ref}^{link}\|_2^2/\sigma_{\text{rot}})$	0.5	1.0
Body link angular velocity	$\exp(-\|\boldsymbol{\omega}_t^{link} - \boldsymbol{\omega}_{ref}^{link}\|_2^2/\sigma_{\text{ang}})$	0.5	1.0
Body link position 3 points	$\exp(-\|\boldsymbol{q}_t^{\text{l3p}} - \boldsymbol{q}_{ref}^{\text{l3p}}\|_2^2/\sigma_{\text{l3p}})$	1.6	0.03
Body link position feet	$\exp\left(-\|\boldsymbol{q}_t^{\text{lf}} - \boldsymbol{q}_{ref}^{\text{lf}}\|_2^2/\sigma_{\text{lf}}\right)$	2.1	0.03
Object position	$\exp(-\|\boldsymbol{q}_t^{o} - \boldsymbol{q}_{ref}^{o}\|_2^2/\sigma_{\text{o\_pos}})$	1.0	0.3
Object velocity	$\exp(-\|\boldsymbol{q}_t^{o} - \dot{\boldsymbol{q}}_t^{o}\|_2^2/\sigma_{\text{o\_vel}})$	1.0	0.3
Object rotation	$\exp(-\|\boldsymbol{\vartheta}_t^{o} \ominus \boldsymbol{\vartheta}_{ref}^{o}\|_2^2/\sigma_{\text{o\_rot}})$	0.5	0.3
Interaction graph	$\exp(-\|s_t^{ig} - s_{ref}^{ig}\|_2^2/\sigma_{\text{o\_ig}})$	2.0	0.3
	Regularization		
Joint position limits	$\mathbb{I}(\boldsymbol{q} \notin [\boldsymbol{q}_{\text{soft-min}},\ \boldsymbol{q}_{\text{soft-max}}])$	-10.0	
Joint velocity limits	$\mathbb{I}(\dot{\boldsymbol{q}} \notin [\dot{\boldsymbol{q}}_{\text{soft-min}},\ \dot{\boldsymbol{q}}_{\text{soft-max}}])$	-5.0	
Joint torque limits	$\mathbb{I}(\boldsymbol{\tau} \notin [\boldsymbol{\tau}_{\text{soft-min}},\ \boldsymbol{\tau}_{\text{soft-max}}])$	-5.0	
Slippage	$\|\dot{\boldsymbol{q}}_{\text{xy}}^{\text{feet}}\|_2^2 \cdot \mathbb{I}[\|F^{\text{feet}}\|_2 \geq 1]$	-1.0	
Torque	$\|\boldsymbol{\tau}\|_2^2$	-1e-6	
Action rate	$\|a_t - a_{t-1}\|_2^2$	-0.5	
Termination	$\mathbb{I}_{\text{termination}}$	-200	

In particular, a penalty term is introduced when the joint positions go beyond the predefined soft limits, where the soft limits are obtained by symmetrically scaling the hard limits with a fixed ratio ($\alpha = 0.95$):

$$m = (\boldsymbol{q}_{\min} + \boldsymbol{q}_{\max})/2, \tag{1}$$

$$d = \boldsymbol{q}_{\max} - \boldsymbol{q}_{\min}, \tag{2}$$

$$\boldsymbol{q}_{\text{soft-min}} = m - 0.5 \cdot d \cdot \alpha, \tag{3}$$

$$\boldsymbol{q}_{\text{soft-max}} = m + 0.5 \cdot d \cdot \alpha, \tag{4}$$

where $\boldsymbol{q}$ is the joint position. The same procedure is applied to compute the soft limits for joint velocity $\dot{\boldsymbol{q}}$ and torque $\boldsymbol{\tau}$. We also project the vector of feet orientation to the x-y plane by $\mathcal{P}(*)$, computing the difference between the forward of body and the orientation of feet to penalize the undesired feet actions.

B Hyperparameter Setting

B.1 Domain Randomization

Since the physical characteristics of the simulator cannot accurately represent the real situation, we incorporate domain randomization (DR) during training to improve the transferability of trained

Table 2: Domain randomization settings.

Term	Coefficient
Large Domain Randomizations	
Friction	$\mathcal{U}(0.5,\ 1.25)$
PD gain	$\mathcal{U}(0.8,\ 1.25)$
Link mass(kg)	$\mathcal{U}(0.8,\ 1.2)$
Object mass(kg)	$\mathcal{U}(0.8,\ 1.2)$
Base CoM offset(m)	$\mathcal{U}(-0.1,\ 0.1)$
Torque	$\mathcal{U}(0.5,\ 1.5)$
Control delay (steps)	$[0, 2]$
Normal Domain Randomizations	
Friction	$\mathcal{U}(0.7,\ 1.2)$
PD gain	$\mathcal{U}(0.7,\ 1.2)$
Link mass(kg)	$\mathcal{U}(0.95,\ 1.05)$
Object mass(kg)	$\mathcal{U}(0.7,\ 1.2)$
Base CoM offset(m)	$\mathcal{U}(-0.1,\ 0.1)$
Torque	$\mathcal{U}(0.7,\ 1.2)$
Control delay (steps)	$[0, 2]$

Table 3: Training Hyperparameters.

PPO	
Hyperparameter	Value
Optimizer	Adam
Training Env	4096
Mini Batches	4
Learning epochs	5
Epoch Horizon	24
Entropy coefficient	0.01
Value loss coefficient	1.0
Clip param	0.2
Max grad norm	1.0
Init noise std	0.8
Learning rate (β)	1e-3
Desired KL	0.01
GAE decay factor(λ)	0.95
GAE discount factor(γ)	0.99
Actor hidden dims	[1024, 512, 256]
Critic hidden dims	[1024, 1024, 128]
MLP Activation	ELU
Soft Actor-Critic	
Hyperparameter	Value
Optimizer	Adam
Batch size	512
Update num	8
Action dim	19
State dim	192
Action range	[-1, 1]
Replay buffer size	4096*20
Init temperature	0.1
Learning rate	3e-4
Actor MLP size	[256, 256]
Critic MLP size	[256, 256]
MLP Activation	ELU

polices to real-world or other simulation settings. We tested two sets of DR: nomral DR and large DR. The specific settings are given in Table 2.

B.2 Reinforcement Learning Hyperparameter

The HOI policy-related PPO hyperparameters and meta policy-related SAC hyperparameters are shown in Table 3.

B.3 InterReal Framework Hyperparameter

In InterReal framework, the number $c_1 = 19$ of selected key body links number on the G1 humanoid robot and the number $c_2 = 1024$ of feature points chosen on the box object. Additionally, we set the number of epochs for stopping decay to $c_4 = 5000$ in $\sigma(t) = clip\left(1 - \frac{c_4}{t}, \delta, 1.\right)$. This means that when PPO is trained after 5000 epochs, the policy scale $\sigma(t) = \delta$, where factor $\delta = 0.1$. We define each $N = 50$ Inner-Loop PPO cycle as a potential subtask $\mathcal{T}_i$, that is, every 50 PPO epochs are regarded as an interaction of the Outer-Loop meta-policy. Our rewards include tracking task-related rewards and penalty-related rewards, totaling items of $K = 19$.

B.4 PPO States

Considering that some HOI features are difficult to obtain in the real world, we design an asymmetric actor-critic module in the PPO architecture to distill away the privileged features in the actor. The specific dimension details are as follows.

- **Actor state dimensions:** The actor's feature inputs s_t^{actor} include the current and 3-step history of the robot's proprioceptive state s_t^h, the task phase variable p and the global position of object $\boldsymbol{q}_t^o$.
- **Critic state dimensions:** The critic's feature input s_t^{crtic} includes extra privileged information, such as the base linear velocity of the robot root, the body position of the reference motion, the difference between the current and reference body link positions, the global object information and the interaction graph information. The details are given in Table 4.

Table 4: Actor and critic state dimensions in the PPO.

Feature term	Actor Dim	Critic Dim
Joint position	23×4	23×4
Joint velocity	23×4	23×4
Root angular velocity	3×4	3×4
Root projected gravity	3×4	3×4
Reference motion phase	1×4	1×4
Actions	23×4	23×4
Global object position	3×1	3×3
Root linear velocity	–	3×4
Global object velocity	–	3×4
Object rotation	–	4×4
Interaction graph	–	189×2
Key robot contact	–	19×1
Reference body link position	–	81
Body link position difference	–	81

Table 5: Actor and critic state dimensions in the SAC.

Feature term	Actor Dim	Critic Dim
PPO rewards	19	19
Tracking errors	5	5
Reference motion phase	1	1
Global object position	3	3
Body link position difference	81	81
Actions	23	23

Table 6: PD controller gains.

Joint name	Stiffness (k_p)	Damping (k_d)
Left/right shoulder pitch	100	2.0
Left/right shoulder roll	100	2.0
Left/right shoulder yaw	50	2.0
Left/right elbow	50	2.0
Waist pitch/roll/yaw	400	5.0
Left/right hip pitch/roll/yaw	100	2.0
Left/right knee	150	4.0
Left/right ankle pitch	40	2.0
Left/right ankle roll	40	2.0
Dex3-1 hands (Box-pushing task)		
Left/right wrist pitch/roll/yaw	8/40/8	0.5/1.5/0.5
index/middle/thumb	2	0.1

B.5 SAC States

The state structure of the SAC algorithm for automatic reward weight learning is shown in Table 5.

B.6 Motion Augmentation

To improve the generalization of HOI motions with respect to object initial positions, we augment each original motion trajectory $\mathcal{M}$ with spatial perturbations applied to the object position. Specifically, a position offset $\Delta \boldsymbol{p}_{xy} = [\Delta x, \Delta y, 0]^\top$ is sampled along the XY-axes of the world coordinate system. We uniformly discretize the range into a 5×5 grid, yielding $c_3 = 25$ offset candidates. The offset ranges are defined as: $\Delta x, \Delta y \in \texttt{linspace}(-0.15, 0.15, 5)$ meters with the bound $\epsilon = 0.15$.

To ensure smooth and physically plausible motion, we apply linear interpolation to the motion 30 frames before and after the interpolation. This guarantees that the augmented trajectories preserve both the hand–object contact constraints and naturalness of the original human motion.

As a result, for each HOI motion $\mathcal{M}$, we obtain a set of 25 augmented variants $\{\mathcal{M}^j\}_{j=1}^{25}$, which are used during training to enhance the robustness and generalization of the learned policy.

C PD CONTROLLER PARAMETER

The gains of the PD controller are listed in Table 6. Based on the experience of previous work, we set the inertia of the ankle joint link to a fixed value of 5×10^{-3} to improve the numerical stability and fidelity of the simulator during training.