TeamHOI: Learning a Unified Policy for Cooperative Human-Object Interactions with Any Team Size

Stefan Lionar[1,2,3] Gim Hee Lee[3]
[1]Garena [2]Sea AI Lab [3]National University of Singapore

https://splionar.github.io/TeamHOI

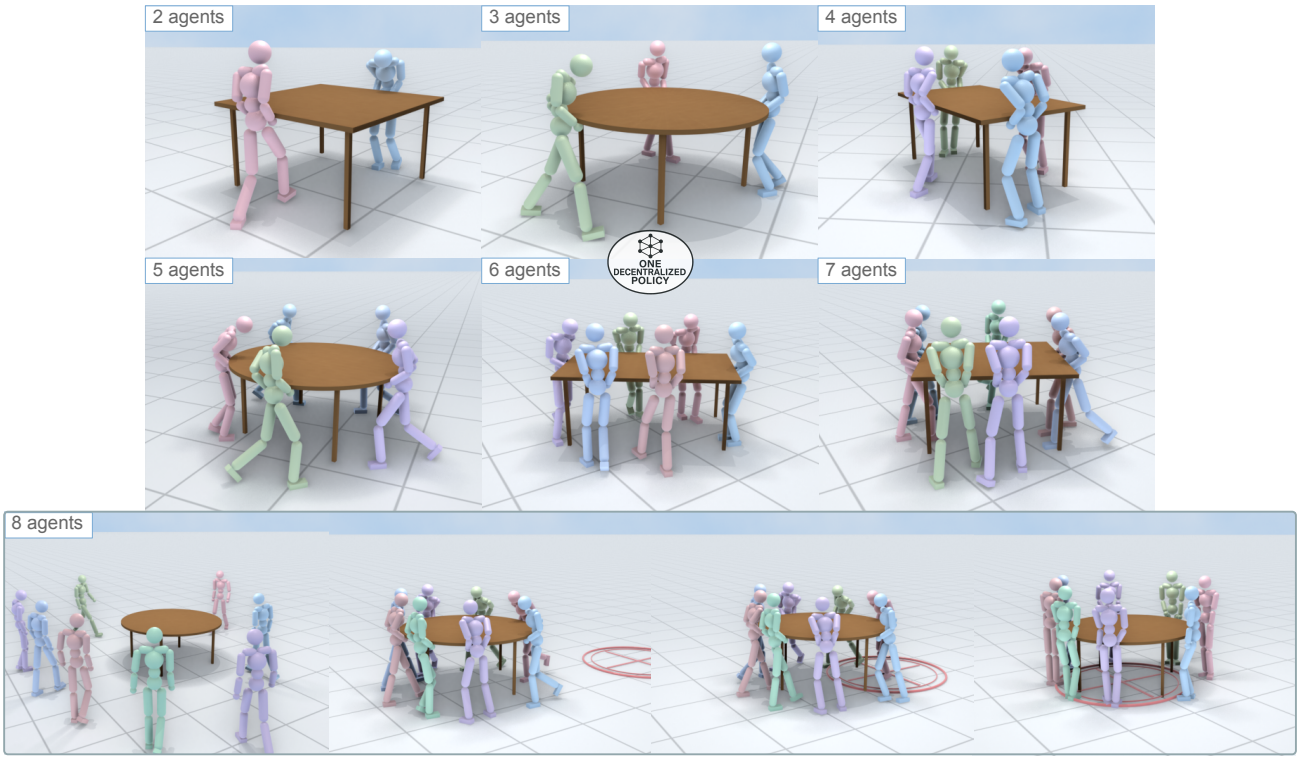

Figure 1. We present **TeamHOI**, a framework for learning a unified decentralized policy for cooperative human-object interactions (HOI) across varying team sizes and object configurations. Our framework enables effective cooperation where each humanoid acts independently from local observations while coordinating with others through a single shared policy. Video demonstrations are provided on our webpage.

Abstract

Physics-based humanoid control has achieved remarkable progress in enabling realistic and high-performing single-agent behaviors, yet extending these capabilities to cooperative human-object interaction (HOI) remains challenging. We present TeamHOI, a framework that enables a single decentralized policy to handle cooperative HOIs across any number of cooperating agents. Each agent operates using local observations while attending to other teammates through a Transformer-based policy network with teammate tokens, allowing scalable coordination across variable team sizes. To enforce motion realism while addressing the scarcity of cooperative HOI data, we further introduce a masked Adversarial Motion Prior (AMP) strategy that uses single-human reference motions while masking object-interacting body parts during training. The masked regions are then guided through task rewards to produce diverse and physically plausible cooperative behaviors. We evaluate TeamHOI on a challenging cooperative carrying task involving two to eight humanoid agents and varied object geometries. Finally, to promote stable carrying, we design a team-size- and shape-agnostic formation reward. TeamHOI achieves high success rates and demonstrates coherent cooperation across diverse configurations with a single policy.

1. Introduction

Physics-based humanoid control and human-object interaction (HOI) have rapidly advanced, enabling virtual humans and robots to walk, grasp, and manipulate objects with realistic motion [10, 52, 61]. Yet many everyday tasks, such as lifting large and heavy items, require multiple agents to coordinate their physical actions. Building humanoids that can cooperate in such settings is a key step toward more capable and intelligent systems. Beyond robotics, this ability also opens up exciting directions for next-generation creative AI applications, such as multi-character animation and interactive game worlds, where virtual humanoids must coordinate naturally with each other.

However, existing physics-based humanoid motion frameworks still face major limitations in both *scalability* and *data diversity* when applied to cooperative HOI. Most approaches rely on fixed-size input MLP policies to generate control actions, and employing such architectures for multi-agent interactions restricts the policy to a fixed team size [37]. Another method omits explicit agent-to-agent communication altogether, and instead relying solely on shared object dynamics as an indirect communication channel [7]. Such designs fail to capture the adaptive nature of real human cooperation, where individuals continuously perceive their teammates' presence and adjust their coordination according to the team's composition and size.

Another key limitation lies in the data source. Many physics-based HOI frameworks leverage the *Adversarial Motion Prior (AMP)* to ensure that learned motions remain natural by regularizing them toward reference motion data. However, reference motion for coordinated multi-human activities is mostly unavailable, necessitating cooperative HOI frameworks to rely on single-human demonstrations. This restriction limits the achievable cooperative behavior. The coordination patterns can only be tied to the motion of one demonstrator, reducing flexibility when handling cooperative HOI with larger groups of agents.

To address these limitations, we propose *TeamHOI*, a framework that enables a single decentralized policy to generalize across any number of cooperative agents. Each agent operates independently using local observations, while sharing the same policy network parameter. To model cooperation with flexible team size, we employ Transformer-based policy network, effectively removing the fixed-size input restriction from MLP policy, and incorporate the states of other agents as *teammate tokens*. The policy is trained in environments instantiated with different team size configurations, exposing it to diverse coordination patterns and allowing it to adapt seamlessly to varying team sizes with their corresponding coordination demands without retraining or fine-tuning.

TeamHOI also addresses the data diversity limitation associated with motion priors. To expand the diversity of feasible cooperative behaviors, we use reference motions from single human actor while masking out the body parts interacting with objects during AMP supervision. We then enforce the masked regions to achieve desirable interactions through task rewards. For example, a sideways walking reference motion can be repurposed for sideways lifting by adapting the hand-object interaction reward. This masking strategy effectively broadens the range of feasible HOI skills, enabling diverse cooperative behaviors to emerge from single-human reference motions.

As a concrete testbed to evaluate our proposed framework, we demonstrate TeamHOI through a challenging cooperative carrying task, as illustrated in Figure 1. In this task, a team of humanoid agents must approach and transport an object, specifically a table that can come in either square, rectangular, or round shape. The agents have to coordinate to establish formations that promote stable lifting, and collectively transport the table to a desired target location. To accomplish this task, we design a *formation reward* that is agnostic to both the table shape and the number of cooperating agents, guiding the agents to distribute themselves into stable positions for carrying. Through extensive experiments, we show that our proposed framework enables a single policy to perform seamlessly across configurations with two to eight cooperating agents, achieving both high success rates and coherent cooperative behaviors.

In summary, our contributions are as follows:

- We introduce *TeamHOI*, a framework that enables a single decentralized policy to perform cooperative human-object interaction with any number of agents.
- We employ a Transformer-based policy network that learns from diverse team-size configurations and adapts the required coordination through teammate tokens.
- We propose a masked AMP strategy that overcomes the data diversity limitation in previous motion-prior methods, expanding the range of feasible and diverse cooperative HOI behaviors.
- We demonstrate the emergent cooperative behaviors of TeamHOI in challenging table-carrying tasks involving varied object shapes and team sizes.
- We design a formation reward that promotes stable carrying, agnostic to both the table shape and the number of cooperating agents.

2. Related Work

2.1. Physics-based Human-Scene Interaction

Physics-based human-scene interaction (HSI) focuses on enabling humanoid agents to interact with objects and environments under realistic physical conditions, including contact, friction, and dynamic constraints. These interactions are typically realized through physics-based control in modern simulators such as Isaac Gym [39] and MuJoCo [59]. A

common training paradigm is reference tracking with deep reinforcement learning (RL) [3, 30, 40, 43, 71, 78], often enhanced by Adversarial Motion Priors (AMP) [44], whose style reward aligns synthesized motions to the reference motions. AMP has proven effective in improving motion realism across diverse downstream tasks [18, 45, 55] and HSI specifically [11, 41, 70]. This line of works has led to a broad range of contact-rich skills, such as in sports environments [28, 29, 37, 62, 64, 75, 76], and everyday object interactions [2, 7, 11, 25, 26, 32, 35, 41, 57, 63, 69, 70, 73, 74].

Beyond mastering individual skills, recent efforts seek broader and more reusable capabilities. Motion-manifold and skill-library approaches learn versatile priors that can be composed across tasks [1, 4, 15, 45, 48], while unified controllers aim to capture diverse behaviors within a single policy [13, 34, 36, 56, 65, 67, 68]. More recently, TokenHSI [42] introduces task tokenization with a Transformer-based policy [60], enabling multi-skill unification for versatile HSI and flexible adaptation to new tasks.

2.2. Multi-Humanoid Interaction and Cooperation

Research on multi-humanoid interactions remains relatively limited compared to single-agent motion synthesis. Several multi-humanoid datasets have been introduced, ranging from everyday human-human activities [6, 21, 22, 33, 50, 72, 77] to choreographic multi-person motions [23, 51]. Building on these data sources, numerous kinematic-based multi-character animation methods have been proposed [5, 8, 9, 16, 27, 49, 53, 54, 66, 80, 81]. Although visually compelling, these approaches rely heavily on high-quality interaction data and cannot guarantee physical plausibility.

To move beyond purely kinematic synthesis, recent works explore physics-based multi-character interactions using kinematic generators [14, 19, 58, 79] and PHC [34], which converts multi-human kinematic motion into state-action pairs for policy learning in physics simulation. This enables interactive physical behaviors among multiple agents [17, 24, 31], albeit without object. Other physics-based efforts focus on crowd navigation [12, 46] that model collision avoidance and group motion.

Another line of works has incorporated object interactions into the multi-humanoid settings. An imitation-based approach [78] demonstrates multi-character interactions involving a shared object, but remains constrained by the scarcity of high-quality multi-human motion capture. Meanwhile, SMPLOlympics [37] showcases multi-humanoid sports behaviors in physics simulation, yet operates with fixed small team sizes. Most relevant to our work is CooHOI [7], which models cooperative human-object interaction by relying on implicit communication through shared object dynamics. However, it does not incorporate the states of other agents, a limitation that is unrealistic for human cooperation where agents continuously perceive and respond to one another. Moreover, its dependence on full-body single-actor reference motions limits the diversity of achievable cooperative behaviors. For instance, the agents are shown to only perform forward and backward lifts and cannot adapt to more varied cooperative HOI strategies.

3. Methodology

Our goal is to develop a unified decentralized policy for cooperative HOI that generalizes across varying team sizes and their corresponding coordination demands. Each agent acts independently based on local observations while incorporating the states of other agents for effective cooperation through a single policy.

3.1. Preliminary

We build on the AMP framework [44], which augments reinforcement learning with motion prior that enforces motion realism. In AMP, a policy π_θ is trained together with a discriminator $D_\phi(s, s')$ that distinguishes short state transitions (s, s') from reference motion data versus those generated by the policy. The discriminator provides a style-based feedback signal that encourages the policy to produce realistic motion transitions.

RL Setup: Each agent observes a proprioceptive state s_t (joint angles, velocities, and root pose), an optional goal state g_t (e.g., object target position), and outputs an action a_t (target joint positions or torques). The environment evolves according to the simulator dynamics to produce the next states and a task reward r_t^{task}. The policy $\pi_\theta(a_t|s_t, s_g)$ is optimized using Proximal Policy Optimization (PPO) [47], maximizing the expected discounted return: $J(\theta) = \mathbb{E}_{\pi_\theta}[\sum_t \gamma^t r_t]$.

Style Reward: To incorporate motion realism, AMP introduces an additional style reward from the discriminator:

$$r_t^{\text{style}} = -\log(1 - D_\phi(s, s')), \tag{1}$$

where (s, s') denotes the current and next states, capturing short-term motion dynamics. The total reward combines both terms as:

$$r_t = r_t^{\text{task}} + \lambda_{\text{AMP}}\, r_t^{\text{style}}, \tag{2}$$

where λ_{AMP} balances task performance and motion realism. The policy is optimized via the PPO objective using r_t, while the discriminator is trained to classify transitions from the policy and those from the reference dataset:

$$\begin{aligned}\mathcal{L}_D = &-\mathbb{E}_{(s,s')^{\text{ref}}}[\log D_\phi(s, s')] \\ &-\mathbb{E}_{(s,s')^{\pi}}[\log(1 - D_\phi(s, s'))].\end{aligned} \tag{3}$$

3.2. TeamHOI Framework

Our TeamHOI framework (Figure 2) reformulates AMP within a flexible multi-agent reinforcement learning setup that scales to an arbitrary number of humanoid agents.

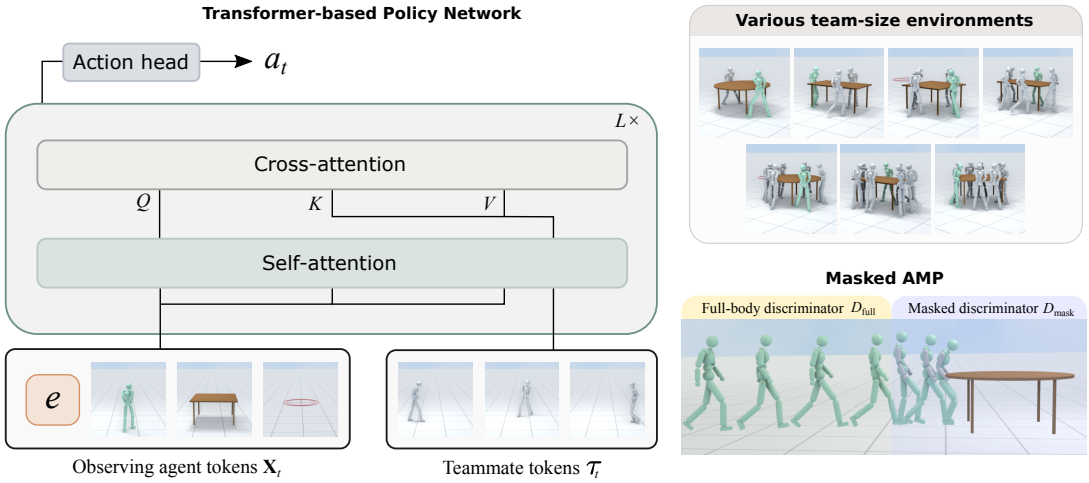

Figure 2. **Overview of TeamHOI framework.** A transformer-based policy network enables coordination between the observing agent (green humanoid) and its teammates (grey humanoids) through alternating self- and cross-attention layers. By training across diverse team-size environments, the framework learns a unified policy that works across different team configurations. To maintain motion realism and enhance skill diversity, a masked AMP strategy blends full-body and masked discriminators based on object interaction.

Policy Network: To enable scalable coordination, we employ a Transformer-based architecture as our policy network, inspired by TokenHSI [42]. Each humanoid agent obtains observation $o_t \triangleq (s_t, g_t)$, which consists of its proprioceptive state s_t and goal states g_t. Each observation component is first processed by a dedicated tokenizer to produce the token sequence of the main observing agent, $\mathbf{X}_t = [e, \mathbf{T}_t^s, \mathbf{T}_t^g]$, where $\mathbf{T}_t^s$ and $\mathbf{T}_t^g$ denote the proprioceptive and goal tokens, and e is a learnable embedding preceding the action head.

To enable coordination with variable team sizes, the observing agent's policy attends to a set of *teammate tokens* $\{\mathcal{T}_t^i\}_{i=1}^{N-1}$, each encoding the cues of another agent (e.g., position, heading direction) expressed in the observing agent's local frame. The transformer backbone consists of L stacks of alternating self-attention and cross-attention layers. Self-attention operates over the observing agent tokens $\mathbf{X}_t$, while cross-attention enables these tokens to attend to the teammate tokens efficiently even when the team size is large. The updated embedding e is passed through an action head to predict control output a_t. The overall policy is defined as $a_t = \pi_\theta(a_t \mid \mathbf{X}_t, \{\mathcal{T}_t^i\}_{i=1}^{N-1})$, where the control output corresponds to the target joint rotation for each actuated degree of freedom for PD controller.

Training a Unified Policy: Within RL framework, each environment represents an independent simulation instance where agents interact with the world, observe states, take actions, and receive rewards. To learn a single policy that generalizes across different collaboration scenarios, we instantiate multiple environments in parallel, each configured with a different number of cooperating agents and their distinctive cooperative rewards. Through this setup, the policy network is trained on diverse multi-agent configurations, gaining exposure to varying interaction dynamics across team sizes. To ensure stable training across mixed team configurations, we normalize PPO advantages separately for each team size. Further details are provided in the supplementary material.

Masked AMP: A major challenge in extending AMP to cooperative HOI is the lack of multi-human reference motion data. Although single-human reference motions can be used, directly regularizing the policy toward them limits the diversity of cooperative behaviors that can emerge, as cooperative tasks often require a wider range of locomotion skills than those present in a single demonstrator.

To address this limitation, we introduce a *Masked AMP* strategy that maintains style realism while allowing diverse HOI behaviors. Specifically, two discriminator networks are trained: one full-body AMP network D_{full} that evaluates complete reference motion, and one masked AMP network D_{mask} that excludes body parts directly interacting with the object (e.g., hands and forearms). During object interaction, the style reward r_t^{mask} is derived from D_{mask}, whereas r_t^{full} from D_{full} is applied when the humanoid is not interacting with the object. The overall blended style reward is:

$$r_t^{\text{style}} = \sigma(\alpha_t)\, r_t^{\text{mask}} + (1 - \sigma(\alpha_t))\, r_t^{\text{full}}, \tag{4}$$

where σ is a sigmoid function operating on continuous interaction indicator α_t (e.g., agent-object distance).

Our formulation shares conceptual similarities with part-wise motion priors (PMP) framework [1], which assembles motion skills from different body segments to enrich physical interaction. However, unlike PMP which learns part-wise priors directly, our method enforces diversity in the masked regions through task rewards to enable adaptable object interactions from limited single-body references.

3.3. Cooperative Carrying Task

As a testbed for our cooperative HOI framework, we design a cooperative carrying task that requires physically grounded coordination among multiple agents. As illustrated in Figure 1, multiple humanoid agents must jointly interact with and transport a large object, specifically a table of varying geometric shapes (e.g., round, rectangular, square). The task has sequential stages that collectively test agents cooperation: 1). *Coordinated formation*: Agents begin at randomized initial locations and must autonomously navigate toward the table. Unlike CooHOI, which assumes oracle-provided per-agent hand target locations, our setup provides no explicit coordination assignment. Instead, 64 candidate contact points are uniformly sampled along the tabletop perimeter on its lower edge, from which agents must infer suitable positions among themselves to ensure stability during lifting. 2). *Cooperative transport*: the agents collectively carry the table toward a designated goal location and then putting it down.

To accomplish this task, our overall reward formulation consists of several components: walking to object, contact, lifting, transport, and put-down—whose detailed expressions are provided in the supplementary.

3.3.1. Formation Reward

Central to the success of this task is how the agents coordinate among themselves to walk into the object while spreading into a formation that promotes stable lifting.

Angular spread reward: To facilitate this, we introduce an *angular spread reward*. It provides a continuous learning signal that promotes the agents to evenly spread themselves around the table, and thus providing a stable support. For each agent, we find its nearest left and right agents and compute the 2D angular gaps to those neighbors about the table's center, denoted as $\Delta\phi_i^{\text{ccw}}$ and $\Delta\phi_i^{\text{cw}}$. For m cooperating agents, the ideal spacing is $2\pi/m$. The reward is then formulated as:

$$r_{\text{ang}} = \exp\Big(-k_\theta \tfrac{1}{2}\big[(\Delta\phi_i^{\text{ccw}} - \tfrac{2\pi}{m})^2 + (\Delta\phi_i^{\text{cw}} - \tfrac{2\pi}{m})^2\big]\Big). \quad (5)$$

Principal-axes coverage reward: Humans most naturally walk forward or backward with symmetric gait, or move sideways through coordinated lateral steps, preferring movement aligned with their local body axes rather than diagonal directions. In cooperative transport, this tendency translates into agents positioning themselves into formations that maximize support along the object's principal axes (its natural axes of rotational stability), and walking along those directions.

The angular spread reward, while effective in promoting balanced coverage for stable lifting, does not enforce these formations. We therefore introduce an additional reward that measures how well the agents' support region spans

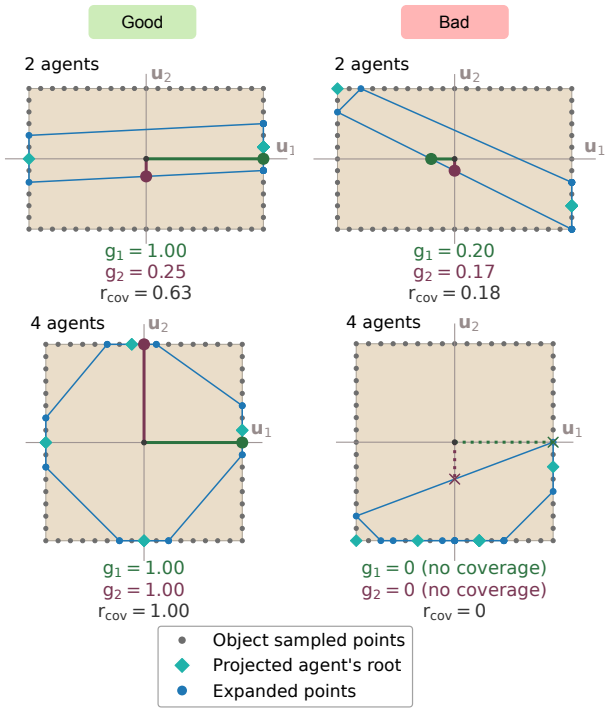

Figure 3. Illustration of our principal-axes coverage reward.

the object's principal axes, as illustrated in Figure 3. First, each agent's root position is projected to the nearest sampled points along the object's perimeter. This set of projected points defines a support polygon via a convex hull. To avoid degenerate case on two agents that can only form a line, each projected point is replaced by two points on its right and left along the perimeter (indices ±2). Let c denote the table's center of mass in the x, y plane. We first compute the principal axes $\mathbf{u}_1, \mathbf{u}_2$ centered on c and measure the distances from c to the support polygon along both directions of each axis, (d_i^+, d_i^-). When the support polygon does not extend across the table center of mass (lying entirely on one side), the distance becomes negative, indicating unbalanced or outside coverage. We therefore clip them as $\tilde{d}_i^{\pm} = \max(0, d_i^{\pm})$. With (ℓ_i^+, ℓ_i^-) denoting the distances from the table center to the table boundary along the same axes, the per-axis coverage is:

$$g_i = \min\left(\frac{\tilde{d}_i^+}{\ell_i^+}, \frac{\tilde{d}_i^-}{\ell_i^-}\right), \quad i \in \{1, 2\}, \quad (6)$$

and the resulting reward is $r_{\text{cov}} = \frac{1}{2}(g_1 + g_2)$. In the supplementary material, we provide a generalized formulation of r_{cov} that supports irregular geometries and non-uniform mass distributions.

r_{cov} and r_{ang} are designed to be complementary. While r_{cov} remains valid under irregular geometries and mass distributions, it can be sparse when agents cluster on one side of the object (yielding zero reward). In contrast, r_{ang} pro-

vides a continuous signal to encourage early dispersion and enable r_{cov} to become non-zero. The final formation reward combines both effects with higher emphasis on r_{cov}:

$$r_{\text{form}} = 0.25\, r_{\text{ang}} + 0.75\, r_{\text{cov}}. \tag{7}$$

4. Experiment

4.1. Implementation Details

Here, we describe the key implementation details of our framework, including the observation states, architecture, and dataset tailored for the cooperative carrying task described in Section 3.3. More implementation details are provided in the supplementary material.

Observation states: Every observing agent receives the following observation components, where each component is expressed in the agent's local coordinate frame:

- **Self-proprioception** $\in \mathbb{R}^{223}$: observing agent's joint states and root kinematics as in standard AMP.
- **Object center** $\in \mathbb{R}^3$: 3D location of the table center.
- **Candidate contact points** $\in \mathbb{R}^{64\times 3}$: 64 uniformly sampled points along the table perimeter on its lower edge. The first index is the point nearest to the agent's root and the remainder are ordered counterclockwise.
- **Nearest hand-to-object points** $\in \mathbb{R}^{2\times 3}$: the candidate contact points nearest to each of the agent's two hands.
- **Target object location** $\in \mathbb{R}^3$: x, y goal of the table center and either the target height z or a binary indicator for lifting vs. putting-down.
- **Teammate cues** $\in \mathbb{R}^{(n-1)\times 9}$: for each of the $n-1$ teammates, includes the x, y root position ($\mathbb{R}^2$), heading direction (6D rotation representation, $\mathbb{R}^6$), and relative angle between the observing agent's root and the teammate's root around the table center in the horizontal plane ($\mathbb{R}^1$). These cues are further encoded as teammate tokens.

Architecture: We use the same observation states and transformer backbone for both the policy and critic networks, differing only in their final outputs. Each observation component is first encoded into a 64-dimensional token using separate three-layer MLP tokenizers with hidden sizes $[256, 128, 64]$. The input tokens include the 64-dimensional learnable embedding e, self-proprioception token, object token (combining object center, candidate contact points, and nearest hand-object points), target-location token, and a variable set of teammate tokens.

The transformer comprises three stacks of alternating self-attention and cross-attention layers, each with two attention heads and a 512-dimensional feed-forward block. The updated embedding e is passed through an MLP with hidden sizes $[1024, 512, 28]$ to predict target joint rotations for PD control in the policy network, and $[1024, 512, 1]$ for critic. Both style discriminators, D_{full} and D_{mask} are implemented as MLPs with hidden sizes $[1024, 512, 1]$ following the standard AMP.

Humanoid and object models: We adopt the Mujoco humanoid model with simplified ball hands without fingers. We design URDF models for three table geometries (square, rectangular, and round) used throughout the cooperative carrying experiments. The table mass ranges from 50 to 70 kg depending on the shape. The square table measures 1.60 m $\times$ 1.60 m, the rectangular table 2.00 m $\times$ 1.20 m, and the round table has a diameter of 2.00 m. The tabletop height is fixed at 0.82 m, slightly below the humanoid's hand position in the default standing pose. This configuration allows our experiment to focus on multi-agent coordination rather than intricate contact interactions.

Reference motions: Our reference motions are from AMASS dataset [38]. Following CooHOI, we adopt 9 walking-related motions from the ACCAD subset and their temporally reversed versions for backward walking, along with 3 sideways-walking motions from the CMU subset. To enable lowering the upper body toward the table and lifting back up, we use 3 pickup motions from the ACCAD subset. These sequences are trimmed before reaching excessively low postures and then reversed to generate the corresponding lifting motions.

4.2. Evaluation

Simulation setup: We train a unified policy for 2-8 agents using our TeamHOI framework. At the start of each episode, agents are randomly initialized on a circle of radius 8 m centered on the object, and the target location is placed in a random direction $[3, 10]$ m away from the table center. Each evaluation episode runs for 600 simulation timesteps.

Baselines: To obtain non-trivial baselines, we substantially adapt the CooHOI [7] framework, denoted as CooHOI*. First, we incorporate our masked AMP strategy to enable diverse realistic locomotion skills for the multi-agent table-carrying. Second, we design a specialized reward function so that any agent, from any initial position, can reach a specified contact point without colliding the table.

Training follows a two-stage pipeline. In the first stage, a single agent is trained to acquire foundational skills: approaching a given contact point, lifting the table, and pushing or dragging it toward the goal. In the second stage, the same policy is extended to multi-agent settings, where coordination emerges implicitly through object dynamics. We train three variants, denoted as CooHOI*-n, where $n \in \{2, 4, 8\}$ indicates the number of cooperating agents. In each configuration, agents are assigned fixed contact points that are manually selected to provide maximum coverage along the object's principal axes. Note that this setup eliminates the need for coordinated formation in the original task, as agents are guided by oracle-defined contact points. Thus, we do not enforce formation reward for CooHOI*, while other rewards are kept the same as ours. Further details of CooHOI* are in the supplementary material.

Table 1. **Quantitative comparison across team sizes (2A, 4A, 8A).** Our method achieves consistently high success rates, collective cooperation, and motion smoothness across all settings using a single unified policy. Unlike CooHOI* baselines, where agent formations are pre-defined, our agents must infer cooperation to establish stable formations autonomously, making the coordination requirement more demanding. Under the heavy-load setting (5× table weights), only our method demonstrates effective cooperation among eight agents. All results are averaged over 10,000 simulation episodes.

Model	**Agents Formation**	**SR (%)↑ / d (m)↓**			t_{coop} **(%)↑**			$\lvert J\rvert$ **(m/s^3) ↓**			**SR$_{5\times}$(%)↑ / $d_{5\times}$ (m)↓**	
		2A	4A	8A	2A	4A	8A	2A	4A	8A	4A	8A
CooHOI*-2	Pre-defined	97.5 / 0.19	73.2 / 1.49	10.1 / 4.25	90.3	54.6	1.0	48.3	85.0	189.9	0.0 / 6.38	0.4 / 6.00
CooHOI*-4	Pre-defined	95.5 / 0.18	94.5 / 0.36	61.5 / 1.83	**96.0**	92.1	27.2	40.7	38.6	96.7	1.2 / 5.34	14.2 / 4.26
CooHOI*-8	Pre-defined	29.4 / 3.60	52.4 / 2.68	42.2 / 3.83	93.8	93.6	81.6	**39.5**	**36.5**	45.2	0.1 / 6.22	4.1 / 5.73
Ours	Learned coop.	**99.1 / 0.06**	**99.2 / 0.08**	**97.5 / 0.18**	95.2	**96.1**	**90.1**	51.0	44.7	**34.2**	**3.5 / 4.78**	**81.1 / 0.49**

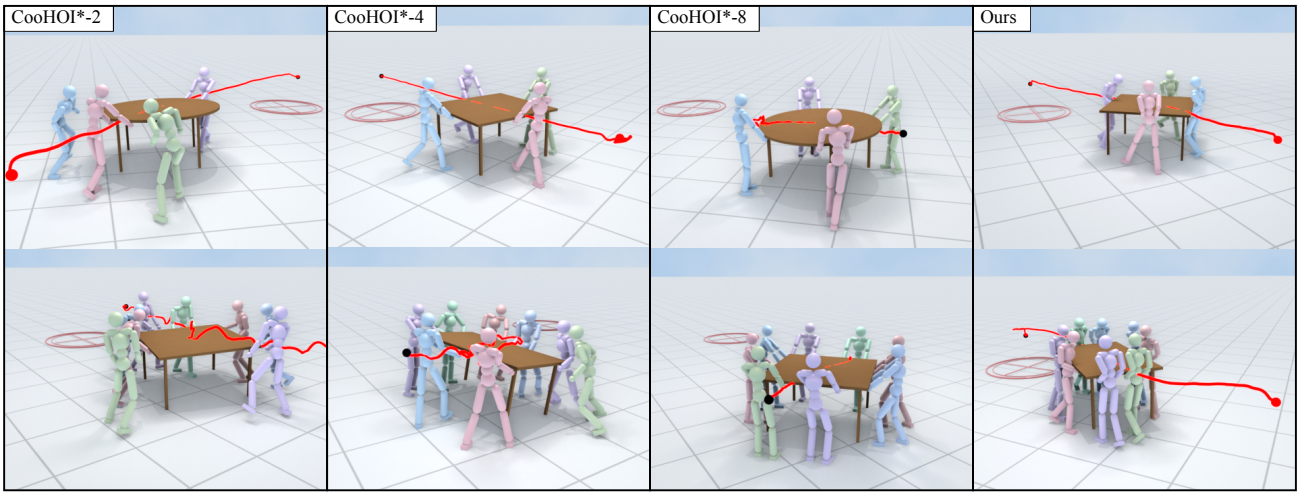

Figure 4. **Qualitative comparison across 4-agent (top) and 8-agent (bottom) configurations.** Our method produces synchronized and stable teamwork across both cases, whereas the CooHOI* baselines exhibit limited or ineffective cooperation. Red line indicates the table's movement trajectory, and the black dot marks its final position at the end of each episode.

Metrics: We evaluate the cooperative carrying task using four quantitative metrics that capture task success, cooperation quality, and motion smoothness:

- **Success rate (SR):** Fraction of episodes with successful task completion, where the object-target distance reaches 0.03 m. Higher is better.
- **Distance to target (d):** Euclidean distance (in meters) between the table center and target location at the end of episode. For successful episodes, d is reported as 0.03 since rewards are not enforced after putdown and baseline models exhibit erratic movements. Lower is better.
- **Cooperative time ratio (t_{coop}):** Fraction of the transport duration during which all agents maintain contact with any of the 64 contact points, indicating consistent collective contribution. Higher is better.
- **Mean absolute jerk ($|J|$):** Average magnitude of the third derivative of the 64 contact point trajectories, capturing transport motion smoothness. Lower is better.

Results: We test each model on the cooperative carrying task with 2, 4 and 8 cooperating agents. Table 1 presents the quantitative results averaged over 10,000 simulations for each cooperative scenario. Our method consistently achieves high success rates, collective cooperation, and smooth motion across all configurations, demonstrating the effectiveness of a single unified policy obtained from our framework. In contrast, the CooHOI* baselines exhibit strong dependence on the specific team size they are trained for. CooHOI*-2 performs well only for two agents but fails to exhibit coordinated behaviors when scaled to larger teams. Similarly, CooHOI*-4 maintains good performance up to four agents but deteriorates sharply beyond that configuration, while CooHOI*-8 struggles to establish effective coordination even within its own setup. We further evaluate the models under a heavy-load setting (5× table weight). The task becomes too challenging for smaller teams, which are barely capable of lifting the table. Notably, only our method demonstrates meaningful cooperation among eight agents, which collectively handle the increased load and achieve a high success rate.

The qualitative results in Figure 4 further highlight the

behavioral differences between our approach and CooHOI* baselines. CooHOI*-2 exhibits competing behaviors between agent pairs, where each pair attempts to move the table independently, resulting in uneven motion and frequent loss of contact. CooHOI*-4 fails to exhibit synchronized cooperation in larger teams, leading to unstable movements. CooHOI*-8 struggles to coordinate effectively when moving the table toward the target, often generating conflicting forces among agents. In contrast, our method achieves globally coherent motion: all agents lift, stabilize, and transport the object as a unified team. These results demonstrate that scalable cooperation across varying team sizes and coordination demands is achieved with the unified decentralized policy trained in our framework. In the supplementary material, we also present more comprehensive experimental results, including robustness to unseen setups and zero-shot generalization to 16-agent configuration.

4.3. Ablation Study

To better understand the contributions of the masked AMP strategy and the proposed formation reward, we conduct an ablation study to examine the effect of each component.

Masked AMP: As shown in Figure 5, incorporating masked AMP significantly improves the overall success rate of the lifting stage. By masking object-interacting body parts during discriminator updates, the policy learns to control hand-object interactions through task rewards rather than being limited by the over-constrained single-human full-body reference motions. Without masking, conflicting objectives between motion realism and object interaction often lead to failed coordination during lifting. Masked AMP alleviates this issue by enabling greater diversity in hand-object interactions. In the subsequent stages, it allows the policy to learn more varied coordination patterns, such as walking or stepping in different directions while carrying the table, despite relying only on single-agent reference motions.

Formation Reward: Figure 6 compares policies trained with and without the proposed principal-axes coverage reward. When the policy is only trained with angular spread reward, agents do not always distribute themselves along the object's principal axes. During mid-training, this setup often produces object trajectories with excessive rotation around the vertical axis. The policy eventually stabilizes but converges to an unnatural diagonal stepping patterns misaligned with the object's principal axes. These unnatural stepping patterns persist across different scenarios and team sizes.

In contrast, incorporating the principal-axes coverage reward encourages agents to align their formations along the object's natural axes of rotational stability. As a result, the team learns to walk in coordinated directions with natural symmetric gaits and balanced support.

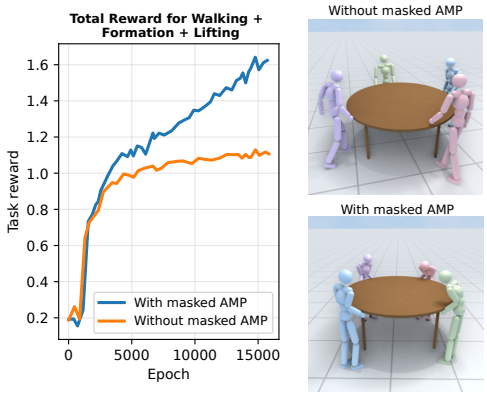

Figure 5. **Ablation on the masked AMP strategy.** Comparison between models trained with and without masked AMP, showing improved task rewards and successful hand-object interactions when masking is applied.

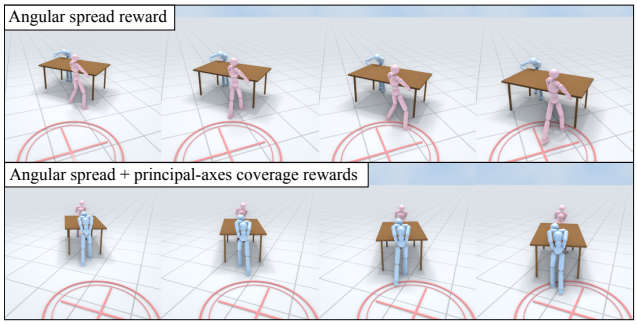

Figure 6. **Formation reward comparison.** Adding principal-axes coverage reward produces stable formations aligned with the object's principal axes, facilitating learned natural locomotion.

5. Conclusion

We present TeamHOI, a unified framework for scalable cooperative human-object interaction that enables a single decentralized policy to generalize across varying team sizes and object configurations. By introducing a Transformer-based policy architecture with teammate tokens, our approach efficiently incorporates teammate cues in a scalable manner, enabling effective inter-agent coordination across varying team sizes. The masked AMP strategy broadens the motion diversity achievable from single-human reference data, while the principal-axes coverage reward encourages stable and natural formations during cooperative transport. Through comprehensive experiments in cooperative carrying task, we demonstrated that TeamHOI achieves coherent, stable, and diverse coordination behaviors across a wide range of multi-agent settings. We believe this work establishes a foundation for scalable, physics-based multi-humanoid control and opens new opportunities for both embodied intelligence and multi-character animation in virtual environments.

Acknowledgment. This research is supported by the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008), and the Tier 2 grant MOET2EP20124-0015 from the Singapore Ministry of Education.

References

[1] Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. Pmp: Learning to physically interact with environments using part-wise motion priors. In *ACM SIGGRAPH*, 2023. 3, 4

[2] Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. Physically plausible full-body hand-object interaction synthesis. In *International Conference on 3D Vision (3DV)*, 2024. 3

[3] Yu-Wei Chao, Jimei Yang, Weifeng Chen, and Jia Deng. Learning to sit: Synthesizing human-chair interactions via hierarchical control. In *AAAI Conference on Artificial Intelligence*, 2021. 3

[4] Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *ACM SIGGRAPH Asia*, 2023. 3

[5] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[6] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[7] Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. Coohoi: Learning cooperative human-object interaction with manipulated object dynamics. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 6, 16

[8] Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. Remos: 3d motion-conditioned reaction synthesis for two-person interactions. In *European conference on computer vision (ECCV)*, 2024. 3

[9] Anindita Ghosh, Bing Zhou, Rishabh Dabral, Jian Wang, Vladislav Golyanik, Christian Theobalt, Philipp Slusallek, and Chuan Guo. Duetgen: Music driven two-person dance generation via hierarchical masked modeling. In *ACM SIGGRAPH*, 2025. 3

[10] Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *arXiv preprint arXiv:2501.02116*, 2025. 2

[11] Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH*, 2023. 3

[12] Brandon Haworth, Glen Berseth, Seonghyeon Moon, Petros Faloutsos, and Mubbasir Kapadia. Deep integration of physical humanoid control and crowd navigation. In *ACM SIGGRAPH*, 2020. 3

[13] Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. Hover: Versatile neural whole-body controller for humanoid robots. In *International Conference on Robotics and Automation (ICRA)*, 2025. 3

[14] Xiaoyu Huang, Takara Truong, Yunbo Zhang, Fangzhou Yu, Jean Pierre Sleiman, Jessica Hodgins, Koushil Sreenath, and Farbod Farshidian. Diffuse-cloc: Guided diffusion for physics-based character look-ahead control. *ACM Transactions on Graphics (TOG)*, 2025. 3

[15] Yiming Huang, Zhiyang Dou, and Lingjie Liu. Modskill: Physical character skill modularization. In *the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[16] Muhammad Gohar Javed, Chuan Guo, Li Cheng, and Xingyu Li. Intermask: 3d human interaction generation via collaborative masked modeling. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[17] Kaiyang Ji, Ye Shi, Zichen Jin, Kangyi Chen, Lan Xu, Yuexin Ma, Jingyi Yu, and Jingya Wang. Towards immersive human-x interaction: A real-time framework for physically plausible motion synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[18] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. In *ACM SIGGRAPH Asia*, 2022. 3

[19] Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015. 16

[21] Kaen Kogashi, Anoop Cherian, and Meng-Yu Jennifer Kuo. Mmhoi: Modeling complex 3d multi-human multi-object interactions. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. 3

[22] Jogendra Nath Kundu, Himanshu Buckchash, Priyanka Mandikal, Anirudh Jamkhandi, Venkatesh Babu Radhakrishnan, et al. Cross-conditioned recurrent networks for long-term synthesis of inter-person human motion interactions. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2020. 3

[23] Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. Music-driven group choreography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[24] Bin Li, Ruichi Zhang, Han Liang, Jingyan Zhang, Juze Zhang, Xin Chen, Lan Xu, Jingyi Yu, and Jingya Wang. Interagent: Physics-based multi-agent command execution via diffusion on interaction graphs. *arXiv preprint arXiv:2512.07410*, 2025. 3

[25] Jianan Li, Tao Huang, Qingxu Zhu, and Tien-Tsin Wong. Physics-based scene layout generation from human motion. In *ACM SIGGRAPH*, 2024. 3

[26] Yitang Li, Mingxian Lin, Zhuo Lin, Yipeng Deng, Yue Cao, and Li Yi. Learning physics-based full-body human reaching and grasping from brief walking references. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[27] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision (IJCV)*, 2024. 3

[28] Libin Liu and Jessica Hodgins. Learning to schedule control fragments for physics-based characters using deep Q-learning. *ACM Transactions on Graphics (TOG)*, 2017. 3

[29] Libin Liu and Jessica Hodgins. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)*, 2018. 3

[30] Libin Liu, KangKang Yin, and Baining Guo. Improving sampling-based motion control. In *Computer Graphics Forum*, 2015. 3

[31] Yunze Liu, Changxi Chen, Chenjing Ding, and Li Yi. Physreaction: Physically plausible real-time humanoid reaction synthesis via forward dynamics guided 4d imitation. In *ACM International Conference on Multimedia*, 2024. 3

[32] Yun Liu, Bowen Yang, Licheng Zhong, He Wang, and Li Yi. Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. *arXiv preprint arXiv:2412.17730*, 2024. 3

[33] Yun Liu, Chengwen Zhang, Ruofan Xing, Bingda Tang, Bowen Yang, and Li Yi. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[34] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[35] Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[36] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[37] Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, et al. Smplolympics: Sports environments for physically simulated humanoids. *arXiv preprint arXiv:2407.00187*, 2024. 2, 3

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 6

[39] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*, 2021. 2, 13

[40] Josh Merel, Saran Tunyasuvunakool, Arun Ahuja, Yuval Tassa, Leonard Hasenclever, Vu Pham, Tom Erez, Greg Wayne, and Nicolas Heess. Catch & carry: reusable neural controllers for vision-guided whole-body tasks. *ACM Transactions on Graphics (TOG)*, 2020. 3

[41] Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. In *International Conference on 3D Vision (3DV)*, 2024. 3

[42] Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3, 4

[43] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 2018. 3

[44] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 2021. 3

[45] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 2022. 3

[46] Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, and Or Litany. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[47] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3, 13

[48] Agon Serifi, Ruben Grandia, Espen Knoop, Markus Gross, and Moritz Bächer. Vmp: Versatile motion priors for robustly tracking motion on physical characters. In *Computer Graphics Forum*, 2024. 3

[49] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[50] Yijun Shen, Longzhi Yang, Edmond SL Ho, and Hubert PH Shum. Interaction-based human activity comparison. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 2019. 3

[51] Li Siyao, Tianpei Gu, Zhitao Yang, Zhengyu Lin, Ziwei Liu, Henghui Ding, Lei Yang, and Chen Change Loy. Duolando: Follower gpt with off-policy reinforcement learning for dance accompaniment. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[52] Kewei Sui, Anindita Ghosh, Inwoo Hwang, Bing Zhou, Jian Wang, and Chuan Guo. A survey on human interaction motion generation. *International Journal of Computer Vision (IJCV)*, 2026. 2

[53] Wenhui Tan, Boyuan Li, Chuhao Jin, Wenbing Huang, Xiting Wang, and Ruihua Song. Think-then-react: Towards unconstrained human action-to-reaction generation. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[54] Mikihiro Tanaka and Kent Fujiwara. Role-aware interaction generation from textual description. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[55] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH*, 2023. 3

[56] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 2024. 3

[57] Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. Maskedmanipulator: Versatile whole-body manipulation. In *ACM SIGGRAPH Asia*, 2025. 3

[58] Guy Tevet, Sigal Raab, Setareh Cohan, Daniele Reda, Zhengyi Luo, Xue Bin Peng, Amit H Bermano, and Michiel van de Panne. Closd: Closing the loop between simulation and diffusion for multi-task character control. In *International Conference on Learning Representations (ICLR)*, 2025. 3

[59] Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *International Conference on Intelligent Robots and Systems (IROS)*, 2012. 2

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[61] Huayi Wang, Wentao Zhang, Runyi Yu, Tao Huang, Junli Ren, Feiyu Jia, Zirui Wang, Xiaojie Niu, Xiao Chen, Jiahe Chen, et al. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*, 2025. 2

[62] Jiashun Wang, Jessica Hodgins, and Jungdam Won. Strategy and skill learning for physics-based table tennis animation. In *ACM SIGGRAPH*, 2024. 3

[63] Wenjia Wang, Liang Pan, Zhiyang Dou, Zhouyingcheng Liao, Yuke Lou, Lei Yang, Jingbo Wang, and Taku Komura. Sims: Simulating human-scene interactions with real world script planning. *arXiv preprint arXiv:2411.19921*, 2024. 3

[64] Yinhuai Wang, Qihan Zhao, Runyi Yu, Hok Wai Tsui, Ailing Zeng, Jing Lin, Zhengyi Luo, Jiwen Yu, Xiu Li, Qifeng Chen, et al. Skillmimic: Learning basketball interaction skills from demonstrations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[65] Jungdam Won and Jehee Lee. Learning body shape variation in physics-based characters. *ACM Transactions on Graphics (TOG)*, 2019. 3

[66] Jungdam Won, Kyungho Lee, Carol O'Sullivan, Jessica K Hodgins, and Jehee Lee. Generating and ranking diverse multi-character interactions. *ACM Transactions on Graphics (TOG)*, 2014. 3

[67] Jungdam Won, Deepak Gopinath, and Jessica Hodgins. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 2020. 3

[68] Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. Uniphys: Unified planner and controller with diffusion for flexible physics-based character control. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[69] Zhen Wu, Jiaman Li, Pei Xu, and C Karen Liu. Human-object interaction from human-level instructions. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

[70] Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified human-scene interaction via prompted chain-of-contacts. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[71] Zhaoming Xie, Jonathan Tseng, Sebastian Starke, Michiel van de Panne, and C Karen Liu. Hierarchical planning and control for box loco-manipulation. *ACM Computer Graphics and Interactive Techniques*, 2023. 3

[72] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[73] Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 3

[74] Sirui Xu, Samuel Schulter, Morteza Ziyadi, Xialin He, Xiaohan Fei, Yu-Xiong Wang, and Liangyan Gui. Interprior: Scaling generative control for physics-based human-object interactions. *arXiv preprint arXiv:2602.06035*, 2026. 3

[75] Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. In *ACM SIGGRAPH*, 2025. 3

[76] Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. Learning physically simulated tennis skills from broadcast videos. *ACM Transactions on Graphics (TOG)*, 2023. 3

[77] Juze Zhang, Jingyan Zhang, Zining Song, Zhanhe Shi, Chengfeng Zhao, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Hoi-mˆ 3: Capture multiple humans and objects interaction within contextual environment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[78] Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH*, 2023. 3

[79] Zihan Zhang, Richard Liu, Rana Hanocka, and Kfir Aberman. Tedi: Temporally-entangled diffusion for long-term motion synthesis. In *ACM SIGGRAPH*, 2024. 3

[80] Zihang Zhang, Shoulong Zhang, Yan Wang, and Shuai Li. Reactffusion: Physical contact-guided diffusion model for reaction generation. In *ACM International Conference on Multimedia*, 2025. 3

[81] Yiwen Zhao, Yang Wang, Liting Wen, Hengyuan Zhang, and Xingqun Qi. Freedance: Towards harmonic free-number group dance generation via a unified framework. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 3

TeamHOI: Learning a Unified Policy for Cooperative Human-Object Interactions with Any Team Size

Supplementary Material

6. Training with Various Team Sizes

6.1. Team-Size Advantage Normalization

PPO [47] algorithm computes the advantage term A_t, which measures how much better an action performs relative to the expected return of the policy's actions in the same state, as estimated by the critic network. The advantages are computed from trajectories collected over a finite time horizon, which determines how far into the future rewards are accumulated. Because their scale can vary across trajectories, the advantages are typically normalized across a batch of trajectories to stabilize training, given as:

$$A_t \leftarrow \frac{A_t - \mu(A)}{\sigma(A) + \epsilon},$$

where $\mu(A)$ and $\sigma(A)$ are the mean and standard deviation computed over the batch.

In our framework, training batches can include data from teams of different sizes, each producing rewards with distinct scales and variances. Normalizing all advantages together across such heterogeneous data can distort their relative magnitudes and influence the accuracy of the policy update signal. Thus, we normalize advantages separately for each team size n:

$$A_t^{(n)} \leftarrow \frac{A_t^{(n)} - \mu_n(A)}{\sigma_n(A) + \epsilon}.$$

As seen in Figure 7, the team-size advantage normalization results in higher task reward.

6.2. Environment Instantiation

We use IsaacGym [39] simulator to train our model. A current limitation of IsaacGym is that each environment in the parallel training must contain the same number of actors, including the humanoid agents and objects. To address this limitation and enable the any team-size unified policy training, we add a dummy ceiling plane in each environment and instantiate a fixed number of agents N. For any environment that requires a smaller team size n, we place the remaining $N - n$ agents on the dummy ceiling. These agents are ignored for the reward calculation, observation states, and gradient computation. Additionally, their PD controllers are disabled. See Figure 8 for an illustration.

7. Reward Functions

Here, we detail all reward components used in the cooperative carrying task, excluding the formation reward r_{form}, angular spread reward r_{ang}, and principal-axes coverage reward r_{cov} already described in the main paper.

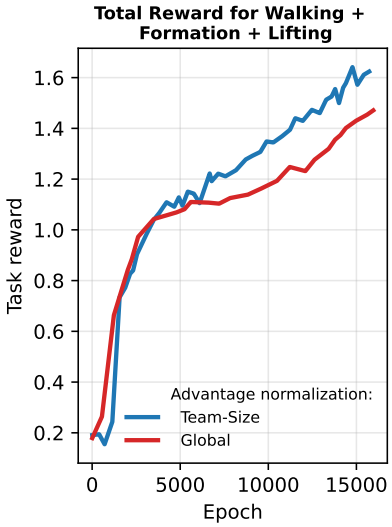

Figure 7. Comparison of task reward curves for models trained with team-size and global advantage normalizations.

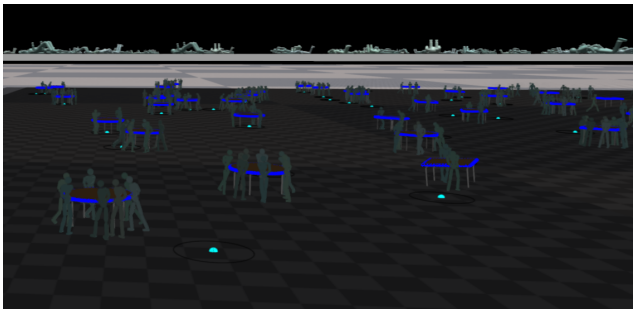

Figure 8. Environment setup in IsaacGym using a dummy ceiling to support flexible team-size training. Extra agents are moved to the ceiling and excluded from observations, rewards, and gradient updates.

7.1. Walking Toward Object

After initialization, each humanoid starts at some distance from the table and is encouraged to walk to the object before the lifting phase. The walking reward is decomposed into three terms that shape the position, velocity, and facing of each agent.

Position: For each agent, we locate the nearest sampled point along the table perimeter, denoted as $\mathbf{p}^*$, and compute the distance to the agent's root $\mathbf{x}_{\text{root}}$ in x-y plane,

$d = \|\mathbf{x}_{\text{root}} - \mathbf{p}^*\|_2$. The agent is encouraged to stand at a target gap $d_{\text{gap}} = 0.3$ m from $\mathbf{p}^*$ by penalizing the squared deviation $\Delta_{\text{gap}} = (d - d_{\text{gap}})^2$. The position reward is:

$$r_{\text{walk}}^{\text{pos}} = \begin{cases} \exp\big(-2.0\Delta_{\text{gap}}\big), & \Delta_{\text{gap}} > 0.04 \text{ m}, \\ 1, & \Delta_{\text{gap}} \le 0.04 \text{ m}. \end{cases} \tag{8}$$

This term acts as an attractive potential that pulls each agent toward the table. It must be balanced by the formation reward r_{form} to ensure that agents spread out while still converging to their appropriate standing regions before lifting.

Velocity: Let $\mathbf{v} \in \mathbb{R}^2$ be the x, y root velocity. We define a desired walking direction $\mathbf{u}^* \in \mathbb{R}^2$ as the inward unit normal in x-y plane associated with the nearest perimeter point $\mathbf{p}^*$. The agent's directional speed s is computed by projecting $\mathbf{v}$ onto $\mathbf{u}^*$, $s = \mathbf{u}^{*\top}\mathbf{v}$. We then encourage the agent to move toward the table within a preferred speed range from $s^*_{\text{low}} = 1.5$ m/s to $s^*_{\text{high}} = 2.5$ m/s. The deviation from this range is expressed using ReLU functions, $\delta_{\text{vel}} = \max(0, s^*_{\text{low}} - s) + \max(0, s - s^*_{\text{high}})$. The velocity reward is:

$$r_{\text{walk}}^{\text{vel}} = \begin{cases} 0, & s \le 0, \\ 1, & \Delta_{\text{gap}} \le 0.04 \text{ m}, \\ \exp\big(-2.0\,\delta_{\text{vel}}^2\big), & \text{otherwise}. \end{cases} \tag{9}$$

Facing: We compute the agent's facing direction by extracting the heading component of its root orientation. Let $\mathbf{f} \in \mathbb{R}^2$ be the facing direction in the x-y plane. We define two target facing directions: the inward normal $\mathbf{u}^*$ at $\mathbf{p}^*$ for near-view alignment, and the direction from the agent toward the table center, $\mathbf{c}^* = \frac{\mathbf{p}_{\text{center}} - \mathbf{x}_{\text{root}}}{\|\mathbf{p}_{\text{center}} - \mathbf{x}_{\text{root}}\|_2}$. The facing reward is computed as:

$$r_{\text{walk}}^{\text{face}} = \begin{cases} \max(0, \mathbf{u}^{*\top}\mathbf{f}), & d \le 1.0 \text{ m}, \\ \max(0, \mathbf{c}^{*\top}\mathbf{f}), & d > 1.0 \text{ m}. \end{cases} \tag{10}$$

7.2. Hand Contact Preparation

After the agents are close to the table with $\|\mathbf{x}_{\text{root}} - \mathbf{p}^*\|_2 \le 1.0$ m, agents are encouraged to reach the hands towards to the table contact points and maintain a reasonable hand configuration for lifting.

Hand reaching: For each agent, let $\mathbf{h}_j \in \mathbb{R}^3$, $j \in \{\text{L}, \text{R}\}$, be the left and right hand positions, and $\{\mathbf{q}_k\}_{k=1}^{64}$ the 64 candidate contact points. We first find the nearest contact point for each hand and its distance, $d_j^{\text{hand}} = \min_k \|\mathbf{h}_j - \mathbf{q}_k\|_2$. A proximity term encourages both hands to approach the contact point:

$$r_{\text{prox}} = \frac{1}{2}\sum_j \exp(-5.0 d_j^{\text{hand}}). \tag{11}$$

In addition, we encourage the hands to reach the lower edge of the table rather than drifting onto the tabletop surface. For each hand $j \in \{\text{L}, \text{R}\}$ with position $\mathbf{h}_j \in \mathbb{R}^3$, let $\mathbf{p}_j^* \in \mathbb{R}^3$ be its nearest sampled perimeter point on the table. We define a contact direction $\hat{\mathbf{v}}_j = \frac{\mathbf{h}_j - \mathbf{p}_j^*}{\|\mathbf{h}_j - \mathbf{p}_j^*\|_2}$. Let $\mathbf{e}_z = (0, 0, 1)^\top$ be the world-up direction. We compute $\cos\theta_j = \hat{\mathbf{v}}_j^\top \mathbf{e}_z$, which measures how much the hand moves upward relative to its associated contact point. The per-hand vertical alignment score is then defined as:

$$r_{\text{above},j} = \begin{cases} \exp\big(-3.0\cos\theta_j\big), & \cos\theta_j > 0, \\ 1, & \cos\theta_j \le 0. \end{cases} \tag{12}$$

The combined term over both hands is:

$$r_{\text{above}} = \frac{1}{2}\big(r_{\text{above,L}} + r_{\text{above,R}}\big). \tag{13}$$

Hand separation: We also encourage a target horizontal separation between the two hands. First, we compute the horizontal separation in the x-y plane, $d_{\text{hand}} = \|(\mathbf{h}_{\text{L}} - \mathbf{h}_{\text{R}})_{xy}\|_2$. We encourage the hands to remain within a preferred separation interval $d^*_{\text{low}} = 0.4$ m and $d^*_{\text{high}} = 0.6$ m. Deviations from this interval are expressed using ReLU functions, $\delta_{\text{sep}} = \max(0, d^*_{\text{low}} - d_{\text{hand}}) + \max(0, d_{\text{hand}} - d^*_{\text{high}})$. We then obtain a hand separation reward:

$$r_{\text{sep}} = \exp\big(-5.0\delta_{\text{sep}}^2\big). \tag{14}$$

To encourage consistent lifting, we penalize vertical mismatch between the two hands. Let z_{L} and z_{R} be their heights, and the reward is defined as:

$$r_{\text{same-z}} = \exp\big(-20.0(z_{\text{L}} - z_{\text{R}})^2\big). \tag{15}$$

Combined reward: The combined hand preparation reward is:

$$r_{\text{hand}} = r_{\text{prox}} \times r_{\text{above}} \times r_{\text{sep}} \times r_{\text{same-z}}, \tag{16}$$

which requires all four terms to be satisfied simultaneously.

7.3. Contact and Lifting

Once the hands are placed near the table edge, additional rewards are activated so that the agents establish contact and lift the table to a desired height.

Contact activation: Let d_j^{hand} be the nearest hand-to-contact distance defined earlier. A per-hand contact score is computed as $\gamma_j = \max\Big(0, 1 - \frac{d_j^{\text{hand}}}{0.06 \text{ m}}\Big)$. We then define a contact reward as the minimum of the per-hand contact scores across the two hands:

$$r_{\text{contact}} = \min(\gamma_{\text{L}}, \gamma_{\text{R}}), \tag{17}$$

and contact indicator for each hand:

$$m_j = \begin{cases} 1, & d_j^{\text{hand}} < 0.04 \text{ m}, \\ 0, & \text{otherwise}, \end{cases}$$

which is used to gate the subsequent lifting and transport rewards.

Lifting height: After contact is established, the hands should lift the table to a target height. Let $\hat{z}_j$ be the height of the contact point associated with hand j, and the target lifting height $z^*_{\text{lift}} = 0.94$ m. We obtain a lifting reward for each hand:

$$\rho_j = \exp(-5.0\,|\hat{z}_j - z^*_{\text{lift}}|). \tag{18}$$

Only hands with valid contact contribute. Therefore, the combined lifting reward is given as:

$$r_{\text{lift}} = \frac{1}{2}\big(m_{\text{L}}\rho_{\text{L}} + m_{\text{R}}\rho_{\text{R}}\big). \tag{19}$$

7.4. Collective Transport

Transport: Once all agents establish contact with the table using both hands, they are encouraged to move the object toward a target location collectively. Let $\mathbf{x}_{\text{obj}} \in \mathbb{R}^2$ be the x, y table position and $\mathbf{x}_{\text{tar}} \in \mathbb{R}^2$ the target location. We define define the transport reward as:

$$r_{\text{transport}} = \begin{cases} \exp\Big(-0.15\,\|\mathbf{x}_{\text{tar}} - \mathbf{x}_{\text{obj}}\|_2^2\Big), & m_{\text{L}} = m_{\text{R}} = 1 \\ & \text{for all agents}, \\ 0, & \text{otherwise}. \end{cases} \tag{20}$$

Carrying alignment: While not strictly required for transport, we include a carrying alignment reward that encourages at least one agent to face toward the target direction while carrying. We identify this agent as the agent farthest from the target. Let $\mathbf{f} \in \mathbb{R}^2$ be this agent's facing direction in the x-y plane, and the desired transport direction:

$$\mathbf{u}_{\text{tar}} = \frac{\mathbf{x}_{\text{tar}} - \mathbf{x}_{\text{obj}}}{\|\mathbf{x}_{\text{tar}} - \mathbf{x}_{\text{obj}}\|_2}.$$

We compute the alignment reward:

$$r_{\text{align}} = \begin{cases} \max\big(0,\, \mathbf{u}_{\text{tar}}^\top \mathbf{f}\big), & m_{\text{L}} = m_{\text{R}} = 1 \text{ for all agents} \\ & \text{and } \|\mathbf{x}_{\text{tar}} - \mathbf{x}_{\text{obj}}\|_2 \geq 0.5 \text{ m}, \\ 1, & m_{\text{L}} = m_{\text{R}} = 1 \text{ for all agents} \\ & \text{and } \|\mathbf{x}_{\text{tar}} - \mathbf{x}_{\text{obj}}\|_2 < 0.5 \text{ m}, \\ 0, & \text{otherwise}. \end{cases} \tag{21}$$

Both $r_{\text{transport}}$ and r_{align} are shared across all agents. During transport, when $m_{\text{L}} = m_{\text{R}} = 1$ for all agents, we set $r^{\text{face}}_{\text{walk}} = 1.0$ so that agents can adjust flexible heading directions while carrying the object collectively.

7.5. Putdown

Once the object reaches the target location, agents must putdown the table and release their hands from the table. Thus, we introduce putdown reward once the object reaches target: $\|\mathbf{x}_{\text{tar}} - \mathbf{x}_{\text{obj}}\|_2 < 0.03$ m.

Hand release: Let z_j be the height of hand $j \in \{\text{L}, \text{R}\}$ and $z^*_{\text{put}} = 0.65$ m the target hand height during putdown. Let d_j^{hand} be the nearest hand–table distance defined earlier. We compute the hand-release reward as:

$$r^{\text{release}}_{\text{put}} = \begin{cases} 1, & d^{\text{hand}}_{\text{L}} > 0.07 \text{ m} \\ & \text{and } d^{\text{hand}}_{\text{R}} > 0.07 \text{ m}, \\ \min\limits_{j \in \{\text{L},\text{R}\}} \exp\big(-5.0\,|z_j - z^*_{\text{put}}|\big), & \text{otherwise}. \end{cases} \tag{22}$$

Zero velocity: During putdown, we also encourage agents to stop moving by applying the following reward:

$$r^{\text{vel}}_{\text{put}} = \exp\big(-2\,\|\mathbf{v}\|_2\big), \tag{23}$$

where $\mathbf{v}$ is the agent's x, y root velocity.

Combined reward: The final putdown reward is a weighted combination of the hand–release and zero–velocity terms:

$$r_{\text{put}} = 0.8\, r^{\text{release}}_{\text{put}} \;+\; 0.2\, r^{\text{vel}}_{\text{put}}. \tag{24}$$

7.6. Total Task Reward

The task reward for the cooperative carrying task is aggregated as follows:

$$\begin{aligned} r^{\text{task}} = {} & 0.2\, r^{\text{pos}}_{\text{walk}} + 0.4\, r^{\text{vel}}_{\text{walk}} + 0.2\, \sqrt{r^{\text{face}}_{\text{walk}}\, r_{\text{ang}}} + 0.6\, r_{\text{form}} \\ & + 0.7\,(r_{\text{hand}} r_{\text{cov}}) + 0.7\, r_{\text{contact}} + 0.7\,(r_{\text{lift}} r_{\text{cov}}) \\ & + 1.0\, r_{\text{transport}} + 0.4\, r_{\text{align}} + 1.0\, r_{\text{put}}. \end{aligned} \tag{25}$$

8. Generalized Principal-Axes Coverage Reward

We elaborate the components to obtain the generalized principal-axes coverage reward r_{cov} that supports irregular geometries (including concave shapes such as L-shape) and non-uniform mass distributions.

Center of mass: Let $\mathcal{X} = \{\mathbf{x}_k \in \mathbb{R}^2\}_{k=1}^N$ denote a set of 2D points sampled from the object's x–y plane (e.g., the tabletop surface). Each point may optionally carry a mass weight $w_k > 0$ representing local density. Uniform density corresponds to $w_k = 1$. The planar center of mass is obtained as $\mathbf{c} = \frac{\sum_{k=1}^N w_k \mathbf{x}_k}{\sum_{k=1}^N w_k}$.

Principal-axes: Next, we obtain the principal-axes $\mathbf{u}_1$ and $\mathbf{u}_2$ from the eigenvectors of the real and symmetric object's

planar inertia matrix $\mathbf{I} = \begin{bmatrix} I_{xx} & I_{xy} \\ I_{xy} & I_{yy} \end{bmatrix}$. Let $\tilde{\mathbf{x}}_k = \mathbf{x}_k - \mathbf{c}$ denote the centered coordinates with components $(\tilde{x}_k, \tilde{y}_k)$. The inertia components are computed as $I_{xx} = \sum_k w_k \tilde{y}_k^2$, $I_{yy} = \sum_k w_k \tilde{x}_k^2$, and $I_{xy} = -\sum_k w_k \tilde{x}_k \tilde{y}_k$. We then compute the eigen decomposition of $\mathbf{I}$ and define $\mathbf{u}_1$ as the eigenvector associated with the smallest eigenvalue, and $\mathbf{u}_2$ as the remaining orthonormal eigenvector.

Boundary extents: We compute the boundary extents ℓ_i^+ and ℓ_i^- along each principal axis $\mathbf{u}_i$ in a manner that remains well-defined for irregular and concave geometries. To this end, we compute the convex hull $\mathcal{H}$ of the object boundary in the planar domain. The boundary extents are the maximum distances from the center of mass $\mathbf{c}$ to the convex hull $\mathcal{H}$ along the positive and negative directions of $\mathbf{u}_i$.

9. Additional Implementation Details

9.1. Training Strategy

Training the unified policy directly with up to eight agents is computationally inefficient due to the long-horizon nature of the cooperative carrying task. We therefore adopt a sequential training with different stages that progressively approaches task completion and increases team size, with early termination triggered whenever any agent falls or table topples. All stages below are trained using a single NVIDIA A100 GPU.

Stage 1: We first train environments instantiated with 1-4 agents to acquire core navigation, contact, and lifting behaviors. At this stage, the transport, alignment, and putdown rewards are disabled (i.e., $r_{\text{transport}} = r_{\text{align}} = r_{\text{put}} = 0$). The full-body discriminator D_{full} is supervised using only forward and sideways walking reference motions, while the masked discriminator D_{mask} additionally receives pickup motions. The target-location token is also masked out during this stage. Training runs for approximately 1.5 days with an episode length of 400 timesteps.

Stage 1+2: Next, we continue training with 2-4 agents to proceed with the coordinated transport and putdown. We enable the remaining task components, including $r_{\text{transport}}$, r_{align} and r_{put}, as well as unmasking the target-location token. Backward walking reference motions are added to supervise D_{mask} to improve locomotion diversity while carrying the object. This stage converges in roughly 5 days with an episode length of 600 timesteps.

Fine-tuning with up to 8 agents: Finally, we fine-tune the unified policy in environments instantiated with 2-8 agents to refine coordination patterns and stabilize collective transport for larger teams. This fine-tuning stage takes about 3 days.

9.2. Training hyperparameters

We train all stages using 1024 parallel environments. For PPO, the minibatch size is set to 16384 when training with up to four agents, and reduced to 8192 for training with up to eight agents. For both PPO and AMP updates, observations belonging to deactivated agents (i.e., those placed above the ceiling) are excluded from the minibatches. For AMP, the reference-motion minibatch size is 4096, and the policy-observation minibatch is set to $1.5\times$ 4096, but excluding the deactivated agents.

Unless otherwise noted, all remaining hyperparameters follow CooHOI [7]. We list the key values in Table 2 for completeness.

Table 2. Key training hyperparameters in our experiment.

Hyperparameter	Value
Horizon length	32
Optimizer	Adam [20]
Learning rate	2×10^{-5}
Task reward weight	0.5
Style reward weight	0.5
PPO clip threshold (ϵ)	0.2
Discount factor (γ)	0.99
GAE parameter (λ)	0.95

10. CooHOI* Baseline

Architecture: CooHOI* follows the same Transformer-based backbone as our method for both the policy and the critic, but replaces the cross-attention with self-attention layer without incorporating teammate tokens. This design mimics the original CooHOI formulation where cooperation emerges solely from the shared dynamics of the object.

Approach-angle reward: We design an approach-angle reward to guide each agent toward its designated contact point while avoiding collision with the table. Let $\mathbf{p}_{\text{des}} \in \mathbb{R}^2$ be the x, y coordinate of the designated point. We first calculate the normalized x, y direction from the object to agent's root: $\hat{\mathbf{a}}_o = \frac{\mathbf{x}_{\text{root}} - \mathbf{x}_{\text{obj}}}{\|\mathbf{x}_{\text{root}} - \mathbf{x}_{\text{obj}}\|_2}$, as well as the normalized x, y direction the object to the designated point: $\hat{\mathbf{p}}_o = \frac{\mathbf{p}_{\text{des}} - \mathbf{x}_{\text{obj}}}{\|\mathbf{p}_{\text{des}} - \mathbf{x}_{\text{obj}}\|_2}$. We then calculate the approach-angle reward based on the cosine similarity between the two directions:

$$r_{\text{approach}} = \frac{\hat{\mathbf{a}}_o^\top \hat{\mathbf{p}}_o + 1}{2}. \tag{26}$$

This yields $r_{\text{approach}} = 1$ when the agent is perfectly aligned toward its target contact point ($\theta = 0°$) and $r_{\text{approach}} = 0$ when it faces the opposite direction ($\theta = 180°$), effectively avoiding collision with the table when agents are initialized in random positions.

The aggregated task reward follows the same structure as Equation 25, except that r_{form}, r_{ang}, and r_{cov} are replaced by the approach-angle reward r_{approach}.

Training strategy: We follow a two-stage training procedure as in CooHOI. In the first stage, a single agent is trained to acquire foundational locomotion and manipulation skills, including approaching the table, maneuvering toward the designated contact point, establishing contact, lifting (or tilting) the table to the target height, and subsequently pushing or dragging it toward the goal. To simplify learning, the friction between the table legs and the ground is set to zero and the table mass is reduced by half during this stage. Training runs for approximately 3 days.

Multi-agent cooperation is then introduced in the second stage by resuming from the single-agent checkpoint. Separate models are trained for team sizes of 2, 4, and 8 agents, denoted as CooHOI-2, CooHOI-4, and CooHOI*-8, respectively. CooHOI*-2 converges in roughly 2 days. CooHOI*-4 requires about 6 days, and CooHOI*-8 continues from the 4-agent checkpoint and trains for an additional 5 days. All models are trained with an episode length of 600 timesteps.

Contact point assignment: To reduce inter-agent collisions when spawning large teams, we enforce a consistent geometric mapping between agents and their designated contact points. All contact points are sorted counterclockwise, starting from the bottom-left corner. After agents are initialized, they are indexed in the same counterclockwise order, also starting from the bottom-left position. A one-to-one assignment is then performed between agent indices and contact points following this order.

11. More Experimental Results

Unified policy across all team sizes: To complement the results presented in the main paper, Table 3 reports the full performance of our unified policy across all team sizes from 2 to 8 agents under both normal ($1\times$) and heavy ($5\times$) table weights. Beyond the configurations shown in the main paper (2A, 4A, 8A), we additionally include intermediate team sizes (3A, 5A, 6A, 7A), demonstrating that the same decentralized policy generalizes smoothly across all team sizes without retraining. Under the normal-weight setting, our model consistently achieves near-perfect success rates across all team sizes with consistent cooperation.

Under the heavy-weight setting ($5\times$ table mass), the increased load amplifies the need for coordinated force generation. As team size grows, our unified policy facilitates effective cooperation that leverages the additional mechanical advantage provided by larger groups, resulting in steadily improving success rates with more agents.

Zero-shot generalization: We further evaluate our unified policy under unseen table geometries and team sizes, testing whether the coordinated formation and carrying skills acquired during training transfer to new scenarios. We consider both smaller tables (round with 1.40 m diameter, square 1.30 m $\times$ 1.30 m, rectangular 1.60 m $\times$ 0.90 m) and larger tables (round with 2.40 m diameter, square 2.20 m $\times$ 2.20 m, rectangular 3.0 m $\times$ 1.40 m), all with the same mass density as in the main experiment.

Table 3. Performance of our unified policy across team sizes under normal ($1\times$) and heavy ($5\times$) table weights.

Team size	SR (%) ↑	d (m) ↓	t_{coop} (%) ↑	$\lvert J\rvert$ (m/s^3) ↓
Normal weight ($1\times$)				
2	99.1	0.06	95.2	51.0
3	99.4	0.06	98.3	50.5
4	99.2	0.08	96.1	44.7
5	99.5	0.06	97.3	40.6
6	99.3	0.07	95.9	38.0
7	98.6	0.11	93.7	35.7
8	97.5	0.18	90.1	34.2
Heavy weight ($5\times$)				
4	3.5	4.77	90.9	23.4
5	18.2	2.48	79.0	28.3
6	50.1	1.04	79.0	32.0
7	71.6	0.59	81.2	31.8
8	81.1	0.49	81.5	31.7

Table 4. Performance of our unified policy across team sizes for small and large tables. All results are averaged over 10,000 simulation episodes.

Team size	SR (%) ↑	d (m) ↓	t_{coop} (%) ↑	$\lvert J\rvert$ (m/s^3) ↓
Small tables				
2	93.1	0.37	94.8	63.2
3	97.5	0.14	97.0	64.7
4	98.4	0.12	97.1	55.5
8	96.4	0.23	85.4	45.0
Large tables				
2	71.0	0.85	91.7	53.3
3	82.7	0.46	94.3	52.1
4	85.3	0.51	94.6	48.8
8	84.2	0.93	86.1	45.4
12	80.6	1.14	58.2	45.2
16	74.5	1.41	15.1	46.6

As shown in Table 4, our policy maintains coherent cooperation across all configurations despite this distribution shift. For smaller tables, agents occasionally display slightly stronger lift initiation, resulting in modestly higher jerk, but the transport phase remains stable and success rates stay consistently high. For larger tables, agents still maintain synchronized cooperative behaviors. However, the increased mass and longer moment arms make lifting and stabilizing harder. Consequently, agents can sometimes lose balance, fall, and trigger early termination. The large-table setting is particularly more challenging for two-agent teams, which have less mechanical leverage to stabilize and lift the heavier tables, resulting in slower transport.

We also evaluate zero-shot generalization to 12-agent and 16-agent teams carrying the large tables, pushing the policy far beyond the team sizes encountered during training. The unified policy continues to produce synchronized and coherent motion, achieving relatively high success rates and low jerk, in contrast to the baseline which becomes highly unstable. However, when teams become very large, the tabletop perimeter becomes crowded, and agents have not fully learned to position themselves within tight support gaps, resulting in lower cooperative-time ratios. Nonetheless, our policy exhibits overall robust generalization to object sizes and larger team sizes unseen during training. We provide several qualitative results in Figure 9.

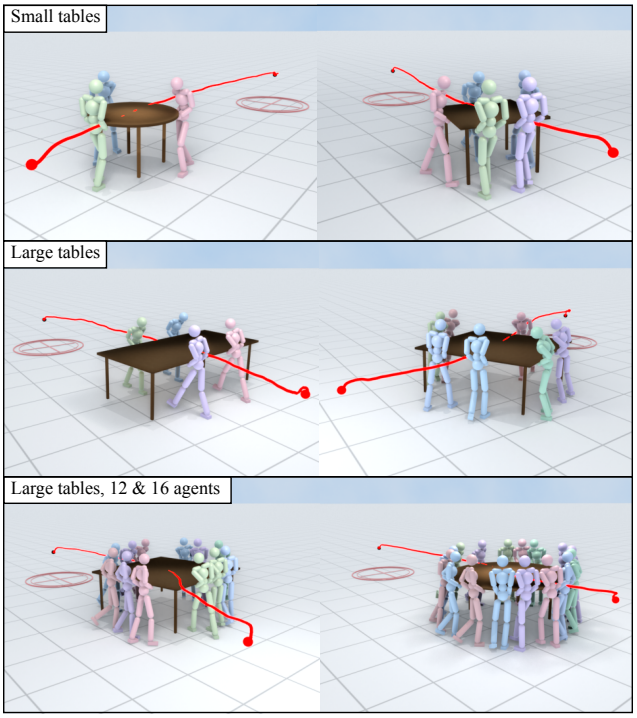

Figure 9. Qualitative visualization of the zero-shot generalization under unseen table geometries and team sizes. Red line indicates the table's movement trajectory, and the black dot marks its final position at the end of each episode.

12. Multiple Affordance Behaviors

While our main experiments focus on a single affordance behavior (edge-lifting), our framework can support multiple affordance behaviors by adapting the task reward. Fig. 10 demonstrates this capability, where agents adapt to either side-holding or edge-lifting depending on their proximity to regions where the corresponding affordances are feasible.

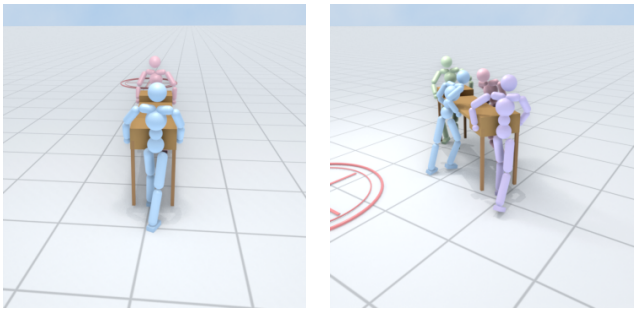

Figure 10. Examples of multiple affordance behaviors learned by adapting the task reward. The agents are able to adapt to side-holding or edge-lifting while being able to walk toward diverse directions. The policy is trained using the same single-human reference motions as our main experiments.