

LHM-Humanoid: Long-Horizon Human Motion Control for Continuous Object Transport in Cluttered Scenes

Haozhuo Zhang¹, Jingkai Sun², Michele Caprio³, Angelo Cangelosi¹, Jian Tang⁴,
Shanghang Zhang⁵, Qiang Zhang^{6*}, and Wei Pan^{7*}

¹ The University of Manchester, Manchester, UK,

haozhuo.zhang@postgrad.manchester.ac.uk, angelo.cangelosi@manchester.ac.uk

² The University of Hong Kong, Hong Kong, China, jingkaisun@connect.hku.hk

³ University of Warwick, Coventry, UK, Michele.Caprio@warwick.ac.uk

⁴ X-Humanoid, jian.tang@x-humanoid.com

⁵ Peking University, Beijing, China, shanghang@pku.edu.cn

⁶ University of Science and Technology of China, Hefei, China, jonyzhang@ustc.edu.cn

⁷ Newcastle University, Newcastle upon Tyne, UK, wei.pan2@newcastle.ac.uk

Abstract. Physics-based human motion control can make a simulated character walk, sit, and manipulate objects with high physical realism. Almost always, though, this happens in short, isolated clips that are re-initialized between interactions. We instead aim for continuous, reset-free long-horizon motion: a physically simulated humanoid that repeatedly walks to a displaced object, lifts it with a balanced whole-body posture, carries it past obstacles, and places it at a goal, over and over within a single uninterrupted take. The hard part is not any individual motion but the transitions between them. Without a reset, each cycle must end in a state that both leaves the object just placed undisturbed and lets the next cycle begin, yet every placement leaves the character off-balance in a non-canonical pose where naive end-to-end reinforcement learning fails. Our key idea is to treat this handoff as a two-sided problem of recoverability: the character must disengage from the object it just placed so the prior success is preserved, and settle into a state from which a balanced continuation exists. Instead of engineering a transition by hand, we learn to shape where each cycle ends so that it lands in this recoverable region. We introduce LHM-Humanoid. One goal-conditioned controller completes a fetch–carry–place cycle and, through a learned release-and-retreat behavior, steers its terminal state into this region; a second controller then takes over from the resulting state distribution. Both are regularized by an adversarial motion prior and distilled into a single goal-conditioned policy that runs the whole sequence as one reset-free rollout. Across 350 cluttered layouts spanning four room types, LHM-Humanoid produces far more successful and stable long-horizon motion than end-to-end RL, hierarchical RL, and prior physics-based human-scene-interaction methods, on both seen and unseen scenes. Trained only on two-object episodes, it extrapolates zero-shot to five-object sequences, where the baselines collapse to near-zero success. The resulting behavior can also be driven by egocentric vision and language. Project page: <https://haozhuo-zhang.github.io/LHM-Humanoid>.

* Corresponding authors.

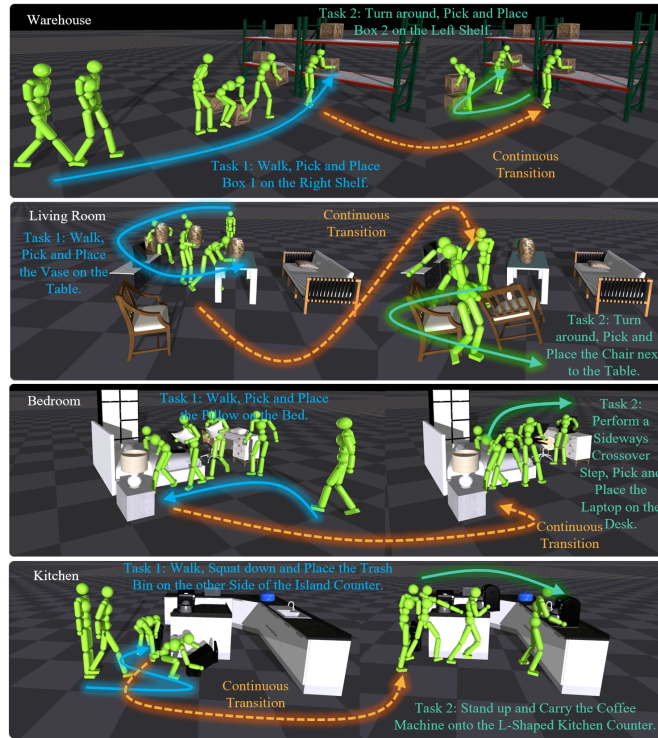


Fig. 1: Overview of LHM. Our approach produces long-horizon whole-body human motion across four distinct environments: Warehouse, Living Room, Bedroom, and Kitchen. A single policy drives a physically simulated character to continuously walk to, pick, carry, and place multiple objects in unseen cluttered scenes, adapting to varied layouts and balance constraints without intermediate resets.

Keywords: Physics-Based Human Motion Control, Whole-Body Human-Scene Interaction, Long-Horizon Motion, Object Transport, Reinforcement Learning

1 Introduction

Generating physically plausible human motion in which a character navigates a scene, interacts with objects, and rearranges its surroundings is a long-standing goal of character animation and embodied simulation. Prior work has made substantial progress in physics-based motion control [22, 27, 34, 50], human-scene interaction [12, 23, 32, 40], and object manipulation [35, 41, 43]. However, most existing settings simplify either the task horizon (to single-step or single-object interactions) or the scene distribution (training and evaluating within a single or fixed scene set) [24, 36, 42].

We push physics-based motion control into a much more demanding regime: continuous, reset-free long-horizon whole-body motion, in which a simulated character transports multiple objects across a cluttered scene. Consider a room where several

objects sit in the wrong place, such as a laptop left on a bed or a trash bin blocking a walkway. Producing believable motion here calls for sustained coordination of locomotion, whole-body manipulation, and object transport over a long horizon. The character must repeatedly handle misplaced objects under balance and obstacle constraints, and recover on its own from the awkward body configurations left after each placement, all in one uninterrupted physically simulated take. Three requirements together separate our setting from prior motion-control work: long-horizon whole-body interaction without resets, cross-scene generalization, and production of the whole sequence from a single unified controller instead of stitched per-skill clips.

Our central observation reframes what makes this hard. The difficulty is not producing any single motion, such as walking, grasping, or placing, but composing them across the seams. In a reset-free episode, the state where one cycle ends becomes the state where the next begins, so stability depends on two things at once: each cycle must end without disturbing the object it just placed, and it must leave the character somewhere a continuation is still feasible. The problem is that every placement leaves the character in a non-canonical, off-balance pose, and naive end-to-end RL breaks down at these transitions. Our answer is a learning principle rather than a scripted maneuver: make the termination distribution recoverable. We train one controller to steer the end of each cycle into the recoverable region, the states from which a balanced continuation provably exists, and a second controller to start from that induced distribution, so the cycles compose without a reset. The hard step is not the release-and-retreat behavior itself, but recognizing recoverability as the governing principle and learning to satisfy it under contact-rich dynamics across hundreds of scenes.

To support research on this problem, we build a dataset of 350 distinct scenes and tasks across four room types, with wide variation in object identities, clutter arrangements, and start/goal configurations. Importantly, it is a task-and-scene benchmark rather than a demonstration dataset: it specifies layouts, object sets, and goals, but provides no scene-specific ground-truth motion. This rules out strict motion mimicry and forces the agent to acquire behavior from task objectives. The large variation within and across room categories also makes room-specific controllers and fixed skill libraries inadequate, so genuine generalization is required.

We introduce LHM-Humanoid, a control pipeline built around a single unified controller. Two goal-conditioned controllers are trained with physics-based RL and an adversarial motion prior [27] for human-likeness. Teacher 1 produces the first fetch-carry-place motion and learns a termination-shaping behavior, realized as release-and-retreat, that drives its terminal state into the recoverable region. Teacher 2 then starts from that induced non-canonical distribution and completes the next cycle. Both are distilled into one unified goal-conditioned controller with DAgger [29], which produces seamless long-horizon motion, including the inter-cycle handoff, as one continuous, reset-free rollout. This controller can be further distilled into a model [7, 43] conditioned on egocentric RGB and natural language.

Contributions:

- **A new motion-control problem and perspective:** continuous, reset-free long-horizon whole-body multi-object transport. We identify that the bottleneck is the inter-cycle handoff, governed by whether each cycle ends in a recoverable state,

rather than the individual motions. In this regime end-to-end RL fails to produce stable motion and prior physics-based methods break down.

- **Learned viability-aware termination shaping:** instead of hand-engineering inter-cycle transitions, we learn a termination behavior (release-and-retreat) that drives each cycle’s terminal distribution into the recoverable region, so the character leaves the object it just placed undisturbed and lands in the set of states from which the next cycle can start. We pair it with a recovery controller trained on exactly that induced distribution, regularize it with an adversarial motion prior, and distill it with DAgger into a single controller.
- **Strong results and a controllable extension:** LHM-Humanoid produces more successful and stable long-horizon interaction than end-to-end RL, hierarchical RL, and prior physics-based human-scene-interaction baselines on seen and unseen scenes. It extrapolates zero-shot from two-object to five-object sequences and can be driven by egocentric RGB and language. We support these results with a diverse testbed of 350 long-horizon cluttered scenes across four room types.

2 Related Work

2.1 Human Motion Modeling and Physically-Based Human-Scene Interaction

Motion synthesis spans kinematic approaches using VAEs, transformers, and diffusion models [3, 13, 15, 16, 19, 20, 48] and physics-based methods that learn controllers under simulation constraints [2, 22, 23, 31, 34, 35, 40, 41, 50]. Physics-based controllers have been advanced through DeepMimic-style imitation [25], adversarial motion priors [27], discrete latent skills [50], reusable skill embeddings [26], language conditioning [17], and expressive control [34].

Human-scene interaction (HSI) further addresses contact-rich behaviors including sitting, lying, and object use, with representative works such as InterPhys [12], InterScene [23], and UniHSI [40]. Earlier physics-based HSI approaches relied on IK, optimal control, or handcrafted controllers [10, 14, 18, 49], while recent methods leverage deep RL [5, 8, 12, 30, 45, 50] to acquire richer skills such as dribbling [36], skateboarding [21], and tool use [44]. Recent scaling efforts including SkillMimic, InterMimic, and TokenHSI [24, 36, 42] further expand physics-based interaction learning, but most controllers remain specialized to narrow task distributions and short, isolated interaction clips. Our work builds on these foundations but synthesizes long-horizon whole-body human-scene interaction as one continuous, reset-free sequence across diverse cluttered scenes.

2.2 Room-Scale Object Rearrangement in Cluttered Scenes

Room-scale rearrangement has been extensively studied in embodied AI and robotics [6, 46], with benchmarks such as Visual Room Rearrangement [38] and OVMM [46], and LLM-based planning approaches [39]. However, these settings typically evaluate kinematic or simplified-embodiment agents and therefore do not model the physics-based whole-body challenges central to our problem—dynamic balance, fall avoidance, whole-body coordination, and contact-rich interaction—so they are not directly comparable

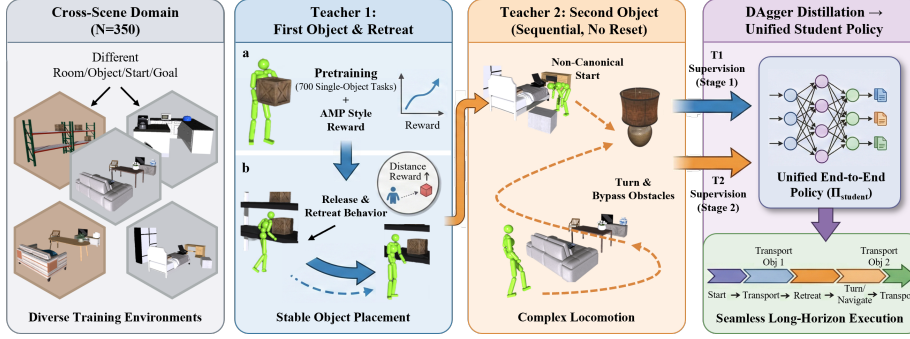


Fig. 2: Overview of the LHM learning framework. The pipeline consists of three stages. **Left:** A diverse dataset of 350 scenes provides training coverage across rooms, objects, and configurations. **Middle:** Two goal-conditioned teacher controllers cover a continuous two-object episode without reset. Teacher 1 completes the first walk–pick–carry–place cycle and executes a release-and-retreat transition to a stable state; Teacher 2 starts from that non-canonical terminal state and completes the next object cycle. **Right:** Both teachers are distilled into one unified goal-conditioned student policy via DAgger, yielding a single network that executes each cycle and the inter-cycle transitions as one continuous, reset-free rollout.

to physically simulated motion control. HumanVLA [43] advances toward heavier object manipulation with a physically simulated human character, but episodes remain relatively short. Our work instead produces long-horizon whole-body object transport as one continuous, reset-free sequence across 350 diverse cluttered scenes; as an extension, we further condition the resulting motion on language and egocentric vision, relating to language-conditioned motion synthesis [4, 15, 47] and vision-language-action policies [7, 9, 37, 43, 51].

3 Method

We formalize reset-free long-horizon whole-body object transport in cluttered scenes and describe our three-stage training pipeline (Figure 2). It is natural to view each fetch-carry-place cycle as a temporally extended action, or option [33],

$$o = \langle \mathcal{I}_o, \pi_o, \beta_o \rangle, \quad (1)$$

with an initiation set $\mathcal{I}_o$, an intra-option policy π_o , and a termination condition β_o . We use this only as an analytical lens: control and reward stay at the primitive timestep, so a full episode can be described as a trajectory over the semi-MDP these cycles induce. What makes the reset-free setting distinctive is that the state where one option ends must fall inside the initiation set of the next. A central part of our design is therefore to shape β_o through release-and-retreat, rather than treat it as a fixed boundary. Because cross-scene generalization is central, all policies are trained jointly over 350 diverse scenes and tasks. There is no scene-specific ground-truth motion, and the configurations vary widely, so fixed skill composition is fragile and direct end-to-end RL on this long-horizon,

contact-rich task tends not to converge. We therefore use a dual-teacher mechanism: Teacher 1 shapes its terminal state into the recoverable region, and Teacher 2 starts from that induced distribution to complete the next cycle. We then apply DAgger [29] to distill both into a single end-to-end student, with a further distillation step producing a vision-language-action (VLA) model for RGB-and-language conditioned control.

3.1 Teacher Policy 1: First-Object Transport with Release-and-Retreat

Teacher 1 completes the first fetch-carry-place cycle. Crucially, it must end inside the recoverable region: it has to leave the object it just placed undisturbed while putting its terminal distribution inside Teacher 2’s initiation set. We therefore treat termination shaping, the requirement that each cycle ends recoverable, as a first-class training objective rather than an afterthought, and realize it through a learned release-and-retreat behavior. If Teacher 1 instead ends off-balance or entangled, Teacher 2 inherits an out-of-distribution state and the error compounds over the rest of the episode. Deciding where Teacher 1 stops is thus what makes the later cycles learnable at all. It also shifts the burden from designing the transition by hand to specifying one property, recoverability of the terminal state, that the policy can then optimize end-to-end.

Pretraining on single-object tasks. We first pretrain Teacher 1 to acquire robust single-object fetch-carry-place behaviors. Each two-object episode is split into two single-object goals, giving 700 single-object tasks that expose the policy to diverse start states, clutter configurations, and spatial layouts. To encourage human-like motion, we add an Adversarial Motion Prior (AMP) [27] style reward r_{style} , computed from a learned discriminator that separates human-like from non-human-like state transitions. We optimize the total reward $r_{\text{total}} = r_{\text{task}} + r_{\text{style}}$, where r_{task} is a shaped task reward for locomotion-to-grasp, grasping, carrying, and placement. The full reward definitions are in the supplementary material.

Release-and-retreat fine-tuning. After pretraining, we fine-tune Teacher 1 to drive its terminal state into the recoverable region. The configurations we target form a recoverable (viable) region [1], the states from which a balanced continuation exists. Concretely, a recoverable terminal state—which also serves as the initiation set from which the second cycle begins—is one where the character has fully released and moved clear of the object it just placed, so that starting the next cycle cannot disturb it, while staying balanced and well-positioned to set off toward the next target. This is closely related to the capturable states in bipedal push recovery [28], and fine-tuning aims to make Teacher 1 end inside this set, within Teacher 2’s domain of attraction. Membership in the viable set is intractable to optimize directly under contact-rich dynamics, so we use release-and-retreat as a tractable surrogate: a reward that grows with the (capped) torso- and hand-to-object distances after placement, and that stays masked until the object is within a small radius of its goal. Rewarding separation only after placement steers the character into this region and holds it there until handoff, rather than merely stepping back. Because the term activates only after a successful placement, it never competes with the task reward during transport, so it does not distort the carrying trajectory; it acts purely as a terminal-state regularizer that reshapes the distribution Teacher 2 must

Algorithm 1 Release-and-Retreat Reward Function

```

1:  $d_{\text{hand2object}} \leftarrow \min(d_{\text{hand2object}}^{\text{left}}, d_{\text{hand2object}}^{\text{right}})$ 
2:  $r_{\text{root2object}} \leftarrow 1 - \exp(-\alpha \cdot d_{\text{root2object}})$ 
3:  $r_{\text{hand2object}} \leftarrow 1 - \exp(-\alpha \cdot d_{\text{hand2object}})$ 
4: if  $d_{\text{root2object}} > \delta_r$  then
5:    $r_{\text{root2object}} \leftarrow 1$ 
6: end if
7: if  $d_{\text{hand2object}} > \delta_h$  then
8:    $r_{\text{hand2object}} \leftarrow 1$ 
9: end if
10: if  $d_{\text{object2goal}} \geq \tau_{\text{object2goal}}$  then
11:    $r_{\text{root2object}} \leftarrow 0$ 
12:    $r_{\text{hand2object}} \leftarrow 0$ 
13: end if
14:  $\text{reward} \leftarrow \beta \cdot r_{\text{root2object}} + (1 - \beta) \cdot r_{\text{hand2object}}$ 

```

handle. Algorithm 1 gives the exact form. The reward grows with the capped torso- and hand-to-object distances after placement (lines 1–9), and stays masked until the object is within a small radius of its goal (lines 10–13), which prevents a premature retreat. The final reward is a weighted combination of the two distance terms (line 14).

3.2 Teacher Policy 2: Recovery Locomotion and Next-Object Transport from Non-Canonical States

Teacher 2 takes over immediately after Teacher 1 completes its release-and-retreat, continuing the episode without any environment reset to complete the next-object fetch-carry-place cycle.

The primary challenge is that the handoff state is non-canonical. After the first placement and retreat, the character may exhibit diverse postures—crouched, leaning, with unevenly spaced feet or displaced hands—and can be oriented arbitrarily, possibly facing away from the next target. The scene may also be partially obstructed by the previously placed object, making subsequent navigation sensitive to incidental contacts.

Teacher 2 therefore solves a compound problem: recovery locomotion to regain a stable walking posture, reorientation toward the next target, obstacle-aware navigation through cluttered passages, and whole-body interaction to grasp, carry, and place the next object. To train it under this broadened distribution, each episode first runs Teacher 1 to completion, then rolls out Teacher 2 from the resulting state and optimizes it with the same goal-conditioned single-object objective used in pretraining. The reward is reused, but the effective training distribution is much harder, because the no-reset setting induces it. This encourages robustness to off-nominal initial states and implicit obstacle-aware navigation without an explicit planner. The design encodes a recovery-from-non-canonical-state mechanism rather than per-object specialization, so it is not tied to the number of objects, as confirmed by zero-shot extrapolation to longer sequences (Sec. 4.2).

3.3 Distillation into a Unified End-to-End Policy

The two teachers are distilled into a single goal-conditioned student that produces every cycle and the transitions between them as one continuous, reset-free rollout. A unified controller is essential here, because it must handle the entire state distribution induced by sequential, contact-rich interaction across diverse scenes. Direct end-to-end RL fails to converge in this setting, due to intractable credit assignment and exploration over long horizons, whereas DAgger-based distillation converges reliably and generalizes across scenes and to longer multi-object settings. Imitation-learning theory predicts this. Naive behavior cloning suffers compounding covariate shift, with a performance gap that grows quadratically in the horizon, $\mathcal{O}(T^2\epsilon)$, whereas aggregating data under the student’s own induced state distribution gives a no-regret bound that is linear in the horizon,

$$J(\pi) \leq J(\pi^*) + uT\epsilon_N + \mathcal{O}(T\delta_N), \quad (2)$$

where ϵ_N is the average online imitation loss, u upper-bounds the cost-to-go difference, and the regret term $\delta_N \rightarrow 0$ as data aggregation proceeds [29]. The linear dependence on T is exactly what a reset-free, long-horizon setting needs, and it matches the gentle degradation of our policy as the number of objects grows (Sec. 4.2).

We use DAgger [29] for distillation. During data aggregation, Teacher 1 supervises the student through the first-object cycle (locomotion-to-grasp, carry, place, release, retreat), and Teacher 2 supervises through the next-object cycle. At each timestep, a finite-state machine (FSM) selects the supervising teacher according to the stage of the episode, which gives stage-appropriate supervision while still training a single unified policy.

3.4 Extension: Distillation into an End-to-End VLA Model

As an extension, we distill the unified student into a vision-language-action (VLA) model conditioned on natural language instructions and egocentric RGB observations. The VLA model replaces privileged oracle state inputs with two perception-grounded modalities: a text instruction specifying the current placement goal, and the character’s first-person RGB stream. Following the same DAgger principle, the unified student provides action supervision, and the VLA model learns to imitate these actions from RGB and language alone. At test time, one can issue consecutive natural language instructions, and the VLA policy produces each fetch-carry-place cycle, with release-and-retreat transitions, within the same episode. This enables long-horizon instruction following with a single end-to-end model. Here the language token is not a redundant goal specifier: under first-person occlusion it is the only cue that tells apart visually similar movable objects (*e.g.*, a pot vs. a coffee-maker on the same stand), and removing it collapses long-horizon success (Sec. 4.3).

4 Experiments

4.1 Dataset and Evaluation Metrics

Our dataset covers four room types—bedroom, living room, kitchen, and warehouse—comprising 350 diverse cluttered scene layouts and long-horizon object-transport

episodes. Scenes contain 79 distinct objects, of which 25 are movable placement targets. Each episode requires completing two consecutive fetch-carry-place cycles in a single continuous rollout (no resets), accompanied by two natural language instructions describing the respective placement goals. To evaluate generalization, we curate 66 unseen tasks with distinct layouts, objects, and instructions as an out-of-distribution benchmark.

Benchmark statistics and diversity. Variation in the benchmark is multi-axis and goes well beyond textures or backgrounds. There are 79 objects (25 movable targets) with large shape dispersion, roughly 4–10 obstacles per scene at varying footprints, heights, and placements, and per-room object initial/goal radii up to about 1.5 m with held-out scene-centroid shifts up to about 0.5 m, on the order of 2–3 character steps. The obstacle layouts are what motivate obstacle-aware carrying and release-and-retreat, rather than straight-line transport. Because the held-out (initial, goal) configurations and centroid shifts are well beyond a single step, success on this split reflects genuine spatial generalization rather than memorization. For human-likeness, AMP uses reference motions from OMOMO [19] and SAMP [11].

Performance is reported via **Success 1** and **Success 2** (per-cycle success rates), **Success All** (both cycles completed in one episode), and **Dist 1/Dist 2** (final placement error in meters). A cycle is counted as successful when the object root position is within 0.2 m of the goal position. Higher success rates and lower distances indicate better performance. Unless otherwise noted, results are averaged over three random seeds, with a typical variance of about ± 3.2 percentage points on Success All.

4.2 Quantitative Results

We report results on the 350 training tasks, 66 unseen tasks, and the VLA extension. **RR** denotes the release-and-retreat objective in Teacher 1; **T** denotes the dual-teacher design; **S** denotes the distilled student.

Baselines. We compare against End-to-End RL, Curriculum RL (the degenerate single-teacher case), Hierarchical RL (five low-level skills—walk, reach-and-grasp, lift, obstacle-aware carry, place-and-release—switched by a privileged FSM), and the physics-based HSI methods HumanVLA [43], InterMimic [42], and TokenHSI [24], alongside our RR/AMP ablations. All methods in Tables 1–4 share identical observation and action interfaces and were re-run with matched environments and compute, so performance differences reflect learning strategy rather than input/output disparities. Exact per-baseline configurations are provided in the supplementary material.

Main results on 350 tasks. Table 1 reveals systematic failure modes across baseline paradigms. End-to-End RL collapses to 0% Success All: credit assignment and exploration are intractable when a policy must discover multi-stage, contact-rich behaviors from sparse rewards over long horizons. Curriculum RL shows that a simple curriculum is not enough, with strong early stages (Success 1: 88.3%) but compounding degradation (Success All: 47.2%). Hierarchical RL and TokenHSI [24] are brittle because their low-level controllers and primitives cover narrow action manifolds, so small changes

Table 1: Results on 350 training tasks.

Method	Success 1 $\uparrow$	Success 2 $\uparrow$	Success All $\uparrow$	Dist 1 (m) $\downarrow$	Dist 2 (m) $\downarrow$
End-to-End RL	2.3%	0.0%	0.0%	2.05	1.97
Curriculum RL	88.3%	48.5%	47.2%	0.26	1.01
Hierarchical RL	37.5%	25.0%	20.8%	0.78	1.13
HumanVLA [43]	42.3%	36.7%	29.9%	0.65	0.97
InterMimic [42]	20.6%	9.8%	7.5%	1.32	1.82
TokenHSI [24]	40.5%	31.5%	27.6%	0.58	1.07
LHM w/o RR	81.1%	60.5%	56.2%	0.52	0.59
LHM w/o AMP	32.9%	14.8%	10.4%	1.09	1.64
LHM-S	87.4%	72.6%	71.1%	0.28	0.50
LHM-T	88.8%	72.9%	72.4%	0.25	0.48

in object pose or clutter create out-of-distribution conditions that compound across stages. InterMimic [42] relies only on task- and scene-aligned supervision, so it transfers poorly to our cluttered benchmark. HumanVLA [43], though unified, degrades over long horizons (Success All: 29.9%) under sequential distribution shift.

Our method addresses these issues with three complementary mechanisms. Release-and-retreat enforces a stable terminal state after each placement, which reduces error propagation; dual-teacher training broadens state coverage beyond canonical trajectories; and DAgger distillation yields a single unified policy that generalizes consistently. LHM-T achieves the best Success All (72.4%) with the lowest placement errors (0.25/0.48 m), and the distilled LHM-S (71.1%) stays close to it. On these same teachers, naive behavior cloning reaches only 32.7% Success All (70.6%/48.3%), while DAgger reaches 71.1% by aggregating labels on the student’s own states. Ablations confirm that both RR (−16.2%) and AMP (−62.0%) are critical. Release-and-retreat pays off only after transport is learned: joint training from scratch reaches a Success 1 of only 10%, against 88.8% after fine-tuning. Given a successful first placement, the next cycle then succeeds in 81.5% of LHM-T episodes, compared with 53.5% for Curriculum RL and 69.3% without release-and-retreat, and the placed object is later displaced in 8.2% of episodes rather than 26.0%.

Results on 66 unseen tasks. On unseen scenes (Table 2), all baselines degrade under distribution shift. Curriculum RL and Hierarchical RL overfit their staged or skill-level training distributions, so novel object poses or clutter push them outside their repertoire and errors compound across the horizon. HumanVLA, InterMimic, and TokenHSI also drop sharply, reflecting limited long-horizon state coverage and brittle motion-only or skill-token supervision. Our method generalizes more robustly. The release-and-retreat objective regularizes the inter-cycle handoff and limits error accumulation after each placement, and the dual-teacher setup exposes the student to a wider distribution of intermediate states than a single canonical trajectory. As a result, the unified policy stays stable under unseen layouts and contact patterns, re-aligns to goals after perturbations, and reaches 63.2% Success All at 0.35/0.50 m placement error. The train→held-out drop for LHM-T is modest, about 7.2/7.8/9.2 points on Success 1/2/All (a relative drop of

Table 2: Results on 66 unseen tasks.

Method	Success 1 $\uparrow$	Success 2 $\uparrow$	Success All $\uparrow$	Dist 1 (m) $\downarrow$	Dist 2 (m) $\downarrow$
End-to-End RL	1.4%	0.0%	0.0%	2.37	2.28
Curriculum RL	75.5%	40.4%	39.5%	0.36	1.11
Hierarchical RL	31.6%	21.8%	17.3%	0.97	1.39
HumanVLA [43]	36.3%	31.3%	25.1%	0.72	1.16
InterMimic [42]	17.2%	8.2%	5.9%	1.47	2.03
TokenHSI [24]	33.9%	26.1%	22.9%	0.67	1.21
LHM w/o RR	69.2%	52.4%	48.6%	0.62	0.60
LHM w/o AMP	27.2%	12.9%	9.0%	1.27	1.86
LHM-S	80.8%	64.9%	61.6%	0.37	0.54
LHM-T	81.6%	65.1%	63.2%	0.35	0.50

roughly 8–13%), and the policy still completes 63.2% of full unseen episodes. Given the substantial spatial variation in Sec. 4.1 and the zero-shot longer-horizon results below, this points to genuine generalization rather than overfitting.

Table 3: VLA extension results (all methods distilled via the same DAgger pipeline).

Method	Success 1 $\uparrow$	Success 2 $\uparrow$	Success All $\uparrow$	Dist 1 (m) $\downarrow$	Dist 2 (m) $\downarrow$
End-to-End RL	1.3%	0.0%	0.0%	2.50	2.49
Curriculum RL	71.5%	41.1%	40.7%	0.37	1.18
Hierarchical RL	29.3%	20.7%	16.5%	1.04	1.45
HumanVLA [43]	34.2%	29.8%	23.5%	0.70	1.26
InterMimic [42]	16.3%	8.4%	6.2%	1.64	2.22
TokenHSI [24]	32.0%	27.0%	23.7%	0.72	1.31
LHM w/o RR	66.7%	52.2%	48.7%	0.60	0.68
LHM w/o AMP	25.3%	14.5%	10.6%	1.39	1.91
LHM-S	76.5%	65.4%	63.7%	0.42	0.56

VLA extension results. Table 3 evaluates all methods under the same DAgger-based distillation into a VLA model, which isolates teacher policy quality as the only variable. Baseline VLA models degrade substantially, because their teachers either fail to acquire robust long-horizon behaviors or rely on narrow skill manifolds that do not survive the modality gap once oracle states are replaced by RGB-and-language inputs. Our VLA model instead keeps strong performance (Success All: 63.7%; Dist 1/2: 0.42/0.56 m) and exceeds the strongest baseline by more than 20 points in Success All (63.7% vs. 40.7%). This shows that the sequential behaviors learned under our framework survive distillation and transfer to egocentric sensing with their spatial precision preserved. Ablations confirm that both RR and AMP are needed here too: removing RR drops Success All to 48.7% and removing AMP to 10.6%.

Table 4: More-than-two-object extension results (models directly evaluated without additional fine-tuning).

Method	Succ1 $\uparrow$	Succ2 $\uparrow$	Succ3 $\uparrow$	Succ4 $\uparrow$	Succ5 $\uparrow$	SuccAll $\uparrow$
End-to-End RL	2.4%	0.0%	0.0%	0.0%	0.0%	0.0%
Curriculum RL	90.4%	51.8%	34.3%	22.3%	6.7%	1.7%
Hierarchical RL	38.6%	27.3%	16.6%	3.8%	0.0%	0.0%
HumanVLA [43]	45.8%	40.5%	21.7%	5.6%	1.2%	0.1%
InterMimic [42]	21.3%	10.2%	2.5%	0.4%	0.0%	0.0%
TokenHSI [24]	41.7%	32.7%	18.9%	4.9%	1.1%	0.1%
LHM w/o RR	82.1%	60.0%	21.6%	4.1%	0.9%	0.0%
LHM w/o AMP	33.6%	14.3%	2.0%	0.9%	0.0%	0.0%
LHM-S	90.8%	76.1%	61.0%	38.5%	20.9%	18.1%

Extension to more than two objects. Table 4 evaluates models trained on two-object episodes and tested directly on longer sequences (up to five objects) without any fine-tuning, on a targeted benchmark of four representative scene types (warehouse, bedroom, living room, kitchen), each instantiated with multiple layouts and several sequential rearrangement tasks (construction details in the supplementary material).

All baselines degrade steeply with episode length and collapse to near-zero Success All beyond three objects. Strong first-object numbers (*e.g.*, Curriculum RL at 90.4%) do not survive the long horizon and fall to single-digit or zero full-episode completion. Our method instead keeps a large advantage at every stage: LHM-S reaches 61.0% on the third object, 38.5% on the fourth, and 18.1% overall on five-object sequences, which no baseline approaches. This gap shows that short-horizon competence does not translate into long-horizon robustness; error propagation and recovery failures dominate as sequences grow. A simple model makes the mechanism concrete; we intend it as an illustration of the trend rather than a derivation. If every cycle is entered from the recoverable region and completed with probability q , an N -object episode succeeds with probability

$$\Pr\left[\text{Succ}_{\text{All}}^{(N)}\right] \approx q^N. \quad (3)$$

When inter-cycle transitions are left unregularized, however, the entry distribution drifts and the per-cycle success degrades geometrically, $q_k \approx q\gamma^{k-1}$ with $\gamma < 1$, giving

$$\Pr\left[\text{Succ}_{\text{All}}^{(N)}\right] \approx \prod_{k=1}^N q\gamma^{k-1} = q^N \gamma^{N(N-1)/2}, \quad (4)$$

an extra $\gamma^{\mathcal{O}(N^2)}$ penalty that mirrors the faster-than-geometric collapse of the baselines beyond three objects. Release-and-retreat keeps γ closer to one than the baselines, by re-entering the recoverable region after each placement, which moves our policy toward the milder q^N regime. This robustness carries over to longer horizons with the same two controllers, with no extra teachers per object, driven by stable inter-stage transitions and broad state coverage rather than task-specific tuning.

4.3 Robustness and Additional Analysis

Robustness of the handoff and source of the gain. The release-and-retreat handoff depends on only a few switching parameters: a speed/settling threshold, a success threshold, and a timing threshold. Sweeping them over a wide range ($\pm 0.5\times$ to $\pm 3\times$ the defaults) changes Success All by at most about ± 2.3 points, so the mechanism does not rely on careful manual tuning (full sweep in the supplementary material). We also verify that the benefit comes from the dual-controller structure rather than from extra state coverage alone. A fully random initial state is not a usable alternative: sampled states include penetrations, falls, and invalid contacts. A single controller trained over a broadened initial-state distribution that already includes the post-placement (non-canonical) configurations reaches only 63.7% Success All (vs. 72.4%), because the transport and recovery rewards interfere under a single value function.

Role of the language modality. Although object/goal state is available during teacher training, the VLA model runs from egocentric RGB and language only. Removing the language token drops Success All from 63.7% to near zero: under first-person occlusion the policy cannot tell apart visually similar movable objects that share a support surface (e.g., a pot vs. a coffee-maker), so language is necessary rather than redundant. The unseen split already includes new language commands, so this also tests instruction generalization.

4.4 Qualitative Visualization

Figure 3 shows rollout snapshots across five diverse environments, where our policy generalizes to novel furniture and obstacle placements without scene-specific fine-tuning. The character performs stable deep-squat lifts and keeps a wide support polygon under heavier loads (warehouse), navigates narrow passages while re-planning body orientation and arm posture to clear obstacles (bedroom and living rooms), and routes objects around counters with clearance instead of straight-line shortcuts (kitchen).

4.5 Failure Analysis

Despite strong overall performance, two failure modes persist. On the 350 training tasks, LHM-T fails the first object in 11.2% of episodes and the second after a successful first placement in 16.4%; within these slices, grasp/navigation shares are 72%/13% and 54%/36%, and the rest is neither. These failures concentrate on flat objects and on narrow, S-shaped passages that are underrepresented in the 350 scenes. Inserting more of these layouts and mining collisions and slipped grasps should cut this tail; the handoff still requires release-and-retreat.

5 Conclusion

We presented LHM-Humanoid, a framework and benchmark for producing continuous, reset-free long-horizon whole-body motion, in which a physically simulated humanoid

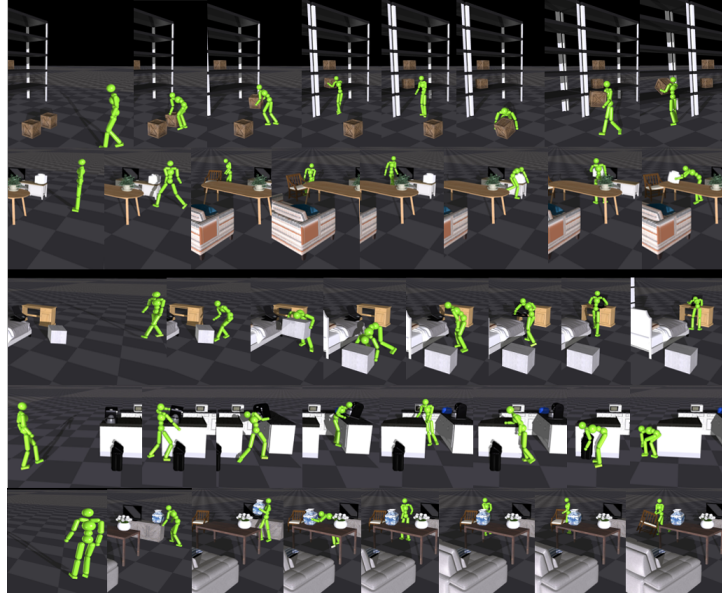


Fig. 3: Qualitative results across diverse unseen scenes. Each row shows a continuous rollout in a different environment (top to bottom: warehouse, living room 1, bedroom, kitchen, living room 2). The policy produces full long-horizon episodes without resets, demonstrating stable whole-body motion, obstacle-aware navigation, and seamless inter-object transitions.

transports multiple objects across cluttered scenes within a single uninterrupted episode. Our key finding is that reset-free composition is governed by recoverability in a two-sided sense: each cycle must end without disturbing the object it just placed, and where the next cycle can begin. Building on this, we integrate learned, viability-aware termination shaping (realized as release-and-retreat), a second controller trained on the induced terminal distribution, and DAGger distillation into one unified policy, which together turn a sequence of contact-rich cycles into a single stable, reset-free rollout. The resulting policy generalizes to unseen scenes and longer sequences and distills into a VLA model, pointing toward a future single recovery-on-demand controller that scales to arbitrarily long horizons without adding controllers.

Limitations The benchmark targets 2–5 objects in physics simulation (Isaac Gym), so longer horizons or transfer to other simulators/morphologies may need adaptation; object and room categories are fixed (no deformables or dynamic obstacles); and the VLA model uses RGB and language only, leaving richer conditioning to future work.

Acknowledgements. This work was partially supported by the ERC Advanced project ETALK (UKRI Funded), the Horizon/UKRI project PRIMI and the US AFOSR project CASPER++.

References

1. Aubin, J.P., Bayen, A.M., Saint-Pierre, P.: Viability theory: new directions. Springer Science & Business Media (2011)
2. Braun, J., Christen, S., Kocabas, M., Aksan, E., Hilliges, O.: Physically plausible full-body hand-object interaction synthesis. In: 2024 International Conference on 3D Vision (3DV). pp. 464–473. IEEE (2024)
3. Cai, Y., Wang, Y., Zhu, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al.: A unified 3d human motion synthesis model via conditional variational auto-encoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11645–11655 (2021)
4. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
5. Cui, J., Liu, T., Liu, N., Yang, Y., Zhu, Y., Huang, S.: Anyskill: Learning open-vocabulary physical skill for interactive agents. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 852–862 (2024)
6. Deitke, M., VanderBilt, E., Herrasti, A., Weihs, L., Ehsani, K., Salvador, J., Han, W., Kolve, E., Kembhavi, A., Mottaghi, R.: Proctor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems* **35**, 5982–5994 (2022)
7. Ding, P., Ma, J., Tong, X., Zou, B., Luo, X., Fan, Y., Wang, T., Lu, H., Mo, P., Liu, J., et al.: Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795* (2025)
8. Dou, Z., Chen, X., Fan, Q., Komura, T., Wang, W.: C-ase: Learning conditional adversarial skill embeddings for physics-based characters. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)
9. Driess, D., Springenberg, J.T., Ichter, B., Yu, L., Li-Bell, A., Pertsch, K., Ren, A.Z., Walke, H., Vuong, Q., Shi, L.X., et al.: Knowledge insulating vision-language-action models: Train fast, run fast, generalize better. *arXiv preprint arXiv:2505.23705* (2025)
10. ElKoura, G., Singh, K.: Handrix: animating the human hand. In: Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation. pp. 110–119 (2003)
11. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11374–11384 (2021)
12. Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–9 (2023)
13. Huang, S., Wang, Z., Li, P., Jia, B., Liu, T., Zhu, Y., Liang, W., Zhu, S.C.: Diffusion-based generation, optimization, and planning in 3d scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16750–16761 (2023)
14. Jain, S., Liu, C.K.: Controlling physics-based characters using soft contacts. In: Proceedings of the 2011 SIGGRAPH Asia Conference. pp. 1–10 (2011)
15. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* **36**, 20067–20079 (2023)
16. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024)
17. Juravsky, J., Guo, Y., Fidler, S., Peng, X.B.: Padl: Language-directed physics-based character control. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)

18. Kim, J., Cordier, F., Magnenat-Thalmann, N.: Neural network-based violinist's hand animation. In: *Proceedings Computer Graphics International 2000*. pp. 37–41. IEEE (2000)
19. Li, J., Wu, J., Liu, C.K.: Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)* **42**(6), 1–11 (2023)
20. Lin, J., Zeng, A., Lu, S., Cai, Y., Zhang, R., Wang, H., Zhang, L.: Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems* **36**, 25268–25280 (2023)
21. Liu, L., Hodgins, J.: Learning to schedule control fragments for physics-based characters using deep q-learning. *ACM Transactions on Graphics (TOG)* **36**(3), 1–14 (2017)
22. Luo, Z., Cao, J., Kitani, K., Xu, W., et al.: Perpetual humanoid control for real-time simulated avatars. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 10895–10904 (2023)
23. Pan, L., Wang, J., Huang, B., Zhang, J., Wang, H., Tang, X., Wang, Y.: Synthesizing physically plausible human motions in 3d scenes. In: *2024 International Conference on 3D Vision (3DV)*. pp. 1498–1507. IEEE (2024)
24. Pan, L., Yang, Z., Dou, Z., Wang, W., Huang, B., Dai, B., Komura, T., Wang, J.: Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5379–5391 (2025)
25. Peng, X.B., Abbeel, P., Levine, S., Van de Panne, M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)* **37**(4), 1–14 (2018)
26. Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)* **41**(4), 1–17 (2022)
27. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)* **40**(4), 1–20 (2021)
28. Pratt, J., Carff, J., Drakunov, S., Goswami, A.: Capture point: A step toward humanoid push recovery. In: *2006 6th IEEE-RAS international conference on humanoid robots*. pp. 200–207. Ieee (2006)
29. Ross, S., Gordon, G., Bagnell, D.: A reduction of imitation learning and structured prediction to no-regret online learning. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. pp. 627–635. JMLR Workshop and Conference Proceedings (2011)
30. Serifi, A., Grandia, R., Knoop, E., Gross, M., Bächer, M.: Vmp: Versatile motion priors for robustly tracking motion on physical characters. In: *Computer graphics forum*. vol. 43, p. e15175. Wiley Online Library (2024)
31. Sferrazza, C., Huang, D.M., Lin, X., Lee, Y., Abbeel, P.: Humanoidbench: Simulated humanoid benchmark for whole-body locomotion and manipulation. *arXiv preprint arXiv:2403.10506* (2024)
32. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. *ACM Transactions on Graphics* **38**(6), 178 (2019)
33. Sutton, R.S., Precup, D., Singh, S.: Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence* **112**(1-2), 181–211 (1999)
34. Tessler, C., Kasten, Y., Guo, Y., Mannor, S., Chechik, G., Peng, X.B.: Calm: Conditional adversarial latent models for directable virtual characters. In: *ACM SIGGRAPH 2023 Conference Proceedings*. pp. 1–9 (2023)
35. Wang, Y., Lin, J., Zeng, A., Luo, Z., Zhang, J., Zhang, L.: Physshoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393* (2023)

36. Wang, Y., Zhao, Q., Yu, R., Zeng, A., Lin, J., Luo, Z., Tsui, H.W., Yu, J., Li, X., Chen, Q., et al.: Skillmimic: Learning reusable basketball skills from demonstrations. *arXiv e-prints* pp. arXiv-2408 (2024)
37. Wang, Y., Li, X., Wang, W., Zhang, J., Li, Y., Chen, Y., Wang, X., Zhang, Z.: Unified vision-language-action model. *arXiv preprint arXiv:2506.19850* (2025)
38. Weihs, L., Deitke, M., Kembhavi, A., Mottaghi, R.: Visual room rearrangement. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 5922–5931 (2021)
39. Wu, J., Antonova, R., Kan, A., Lepert, M., Zeng, A., Song, S., Bohg, J., Rusinkiewicz, S., Funkhouser, T.: Tidybot: Personalized robot assistance with large language models. *Autonomous Robots* **47**(8), 1087–1102 (2023)
40. Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. *arXiv preprint arXiv:2309.07918* (2023)
41. Xie, Z., Tseng, J., Starke, S., van de Panne, M., Liu, C.K.: Hierarchical planning and control for box loco-manipulation. *Proceedings of the ACM on Computer Graphics and Interactive Techniques* **6**(3), 1–18 (2023)
42. Xu, S., Ling, H.Y., Wang, Y.X., Gui, L.Y.: Intermimic: Towards universal whole-body control for physics-based human-object interactions. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 12266–12277 (2025)
43. Xu, X., Zhang, Y., Li, Y.L., Han, L., Lu, C.: Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *Advances in Neural Information Processing Systems* **37**, 18633–18659 (2024)
44. Yang, Z., Yin, K., Liu, L.: Learning to use chopsticks in diverse gripping styles. *ACM Transactions on Graphics (TOG)* **41**(4), 1–17 (2022)
45. Yao, H., Song, Z., Zhou, Y., Ao, T., Chen, B., Liu, L.: Moconvq: Unified physics-based motion control via scalable discrete representations. *ACM Transactions on Graphics (TOG)* **43**(4), 1–21 (2024)
46. Yenamandra, S., Ramachandran, A., Yadav, K., Wang, A., Khanna, M., Gervet, T., Yang, T.Y., Jain, V., Clegg, A.W., Turner, J., et al.: Homerobot: Open-vocabulary mobile manipulation. *arXiv preprint arXiv:2306.11565* (2023)
47. Zhao, K., Wang, S., Zhang, Y., Beeler, T., Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: *European Conference on Computer Vision*. pp. 311–327. Springer (2022)
48. Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 14738–14749 (2023)
49. Zhao, W., Zhang, J., Min, J., Chai, J.: Robust realtime physics-based motion control for human grasping. *ACM Transactions on Graphics (TOG)* **32**(6), 1–12 (2013)
50. Zhu, Q., Zhang, H., Lan, M., Han, L.: Neural categorical priors for physics-based character control. *ACM Transactions on Graphics (TOG)* **42**(6), 1–16 (2023)
51. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: *Conference on Robot Learning*. pp. 2165–2183. PMLR (2023)