

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

Supplementary Material

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

This document provides supplementary details that support—but are not required to understand—the main paper. We collect here the full reward design, implementation and training details, an expanded description of the benchmark and its diversity, exact baseline configurations, and additional robustness analyses. All cross-references of the form “Sec. X”, “Table X”, or “Eq. X” in the main paper that point to the supplementary material refer to the sections below.

A Reward Design and Implementation Details

Task reward. The teacher-policy task reward is a weighted sum of shaped sub-rewards guiding locomotion-to-grasp, grasping, carrying, and placement:

$$r_{\text{task}} = \sum_{i=1}^N \alpha_i \cdot r_i, \quad (\text{S1})$$

where α_i and r_i denote the weight and value of the i -th component, respectively. Table S1 lists all active reward terms and their definitions.

Notation: $\mathbf{p}_r$, $\mathbf{p}_o$, $\mathbf{p}_g$ are the character-root, object, and goal positions; $\mathbf{v}_r$, $\mathbf{v}_o$ are their velocities; $\hat{\mathbf{d}}$ is the unit vector toward the current target; v_t and v_o^{target} are target speeds; $\mathbf{p}_h$ is the character hand position; $\mathbf{p}_j^{\text{pcd}}$ is the object-surface point nearest the hand, not a predefined keypoint; z_o , z_{init} , z_{target} are the object’s current, initial, and desired heights; and $\Delta R(\mathbf{q}_o, \mathbf{q}_g)$ is the rotational error between object and goal orientations.

Style reward. To encourage human-like motion we add an Adversarial Motion Prior (AMP) [3] style reward,

$$r_{\text{style}} = \lambda_{\text{AMP}} \cdot \left[-\log \left(1 - D_{\phi}(s_t, s_{t+1}, a_t) \right) \right], \quad (\text{S2})$$

* Corresponding authors.

Table S1: Active reward terms and their mathematical definitions.

Reward term	Formula
$r_{\text{root2object_vel}}$	$\exp(-2 \cdot (v_t - \mathbf{v}_r \cdot \hat{\mathbf{d}})^2)$
$r_{\text{root2object_pos}}$	$\exp(-0.5 \cdot \ \mathbf{p}_r - \mathbf{p}_o\ ^2)$
$r_{\text{hand2object}}$	$\exp(-5 \cdot \min_j \ \mathbf{p}_h - \mathbf{p}_j^{\text{pcd}}\)$
r_{height}	$\frac{\min(z_o, z_{\text{target}}) - z_{\text{init}}}{z_{\text{target}} - z_{\text{init}}}$
$r_{\text{object2goal_vel}}$	$\exp(-2 \cdot (v_o^{\text{target}} - \mathbf{v}_o \cdot \hat{\mathbf{d}})^2)$
$r_{\text{object2goal_pos_far}}$	$\exp(-1 \cdot \ \mathbf{p}_o - \mathbf{p}_{g,\text{inter}}\)$
$r_{\text{object2goal_pos_near}}$	$\exp(-5 \cdot \ \mathbf{p}_o - \mathbf{p}_g\)$
$r_{\text{object2goal_rot}}$	$\exp(-2 \cdot \Delta R(\mathbf{q}_o, \mathbf{q}_g))$

where D_ϕ is a learned discriminator distinguishing human-like from non-human-like state transitions and λ_{AMP} is a weighting coefficient. The total pretraining reward is $r_{\text{total}} = r_{\text{task}} + r_{\text{style}}$.

Observation and action space. We follow the observation and action space design of HumanVLA [5], comprising proprioception, object/goal state, a waypoint signal, and AMP discriminator observations [3]. The policy outputs PD controller target joint positions. Obstacle avoidance matches HumanVLA: the waypoint signal together with implicit avoidance while carrying.

Training schedule. Teacher 1 is trained for 10,000 epochs; Teacher 2 for 7,000 epochs; the unified student is distilled for 50,000 epochs via DAgger. All training runs in Isaac Gym [1] on two 80 GB NVIDIA A100 GPUs with 16,384 parallel environments. Results are averaged over three random seeds. Simulation runs at 60 Hz; policy inference at 30 Hz. These epoch counts are comparable in scale to prior physics-based synthesis pipelines—*e.g.*, HumanVLA [5] trains 30,000 teacher epochs plus 20,000 DAgger-student epochs at the same 16,384 environments. Because epochs are amortized over 350 distinct scenes, per-scene exposure is modest rather than single-scene over-training, consistent with the small train→held-out gap reported in the main paper.

Network architecture. Both teacher policies follow an AMP-style design with an actor (MLP with ReLU activations) for action production, a critic for value estimation, and a discriminator for the AMP style signal. Running normalization statistics are maintained online. The distilled student is an actor-only network at test time. For the VLA model, a multi-modal actor fuses visual features from an EfficientNet backbone, language features projected by a small MLP (with frozen text features), and normalized proprioceptive features; the fused representation is decoded to control actions via a residual MLP.

B Benchmark Statistics and Diversity

This section expands the dataset summary in the main paper with a per-axis breakdown of the benchmark’s variation. Variation is multi-axis and goes well beyond textures or background assets. (i) Object variation: 79 objects (25 movable targets) with substantial shape dispersion across categories. (ii) Obstacle/layout variation: each scene contains roughly 4–10 obstacles at varying footprints, heights, and placements, which is what motivates the release-and-retreat behavior and obstacle-aware carrying rather than straight-line transport. (iii) Spatial variation: per-room object initial/goal radii reach up to about 1.5 m, with held-out scene-centroid shifts of up to about 0.5 m—on the order of 2–3 character steps (~ 0.75 m each), so the held-out split is not a re-texturing of the training layouts. (iv) Additional variation in object heights, textures, and backgrounds. Because the held-out (initial, goal) configurations and centroid shifts substantially exceed a single step, success on this split reflects genuine spatial generalization rather than memorization.

For the more-than-two-object evaluation in the main paper, we design a targeted benchmark comprising four representative scene types—warehouse, bedroom, living room, and kitchen—each instantiated with five independently configured layouts and at least three distinct sequential object-rearrangement tasks, yielding a diverse yet tractable test bed that probes generalization across environments and task structures.

C Baseline Configurations

We detail the exact configuration of every baseline. The single-teacher variant degenerates to Curriculum RL, as a single teacher trained on canonical trajectories lacks the state coverage needed for long-horizon recovery. Hierarchical RL uses five low-level controllers and a privileged FSM for oracle-based stage switching: walk-to-target, reach-and-grasp, lift, carry with obstacle avoidance, and place-and-release. For HumanVLA [5], we split our 350 long-horizon tasks into 700 short-horizon subtasks matching the HumanVLA training protocol, because the HumanVLA pipeline is designed for single-object transport only; we keep identical reward functions and procedures. InterMimic [4] follows its official pipeline, using the released AMP reference motions, and is then evaluated as a learned policy. TokenHSI [2] adopts the TokenHSI model architecture but is trained on the same five skills as Hierarchical RL.

D Additional Robustness Analysis

Sensitivity of the dual-teacher handoff parameters. The release-and-retreat handoff uses a small number of switching parameters (*e.g.*, a speed/settling threshold s_t , a success threshold succ_t , and a timing threshold t_t). To verify that our framework does not rely on careful manual tuning, we sweep these parameters over a wide range ($\pm 0.5\times$ to $\pm 3\times$ of the default values). As shown in Table S2, all variations remain within about ± 2.3 points of the original Success All (72.4%), indicating that the mechanism is robust to substantial mis-tuning and does not require an extensive search to transfer across configurations.

Table S2: Sensitivity of the dual-teacher switching parameters (excerpt). All values are within ± 2.3 points of Success All of the default setting.

	Succ 1	Succ 2	Succ All	Dist 1	Dist 2
Default	88.8	72.9	72.4	0.247	0.484
$s_t \times 0.5$	88.3	71.2	70.8	0.250	0.502
$s_t \times 3$	87.8	72.1	70.7	0.263	0.504
$\text{succ}_t + 0.5$	88.8	73.1	71.3	0.255	0.461
$t_t - 20$	88.3	70.6	70.1	0.251	0.511

Is the gain from two teachers or merely broader state coverage? A natural alternative is a single controller trained with a broadened initial-state distribution that already covers the post-placement (non-canonical) configurations Teacher 2 encounters. We train such a single controller initialized uniformly over canonical and post-placement states; it reaches only 68.5% Success 2 and 63.7% Success All, versus 72.9% and 72.4% for the dual-teacher design. The two-mode reward landscape (transport vs. recovery) interferes under a single value function, confirming that the gain comes from the dual-controller structure itself rather than from broader state exposure alone.

References

1. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
2. 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)
3. 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)
4. 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)
5. 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)