LUCID: Latent-Skill Unified Control via Imagined Dynamics for Long-Horizon Humanoid Loco-Manipulation

Cheng Guo[1,2*], **Mingzhe Ni**[1], **Angelo Cangelosi**[1], **Arash Ajoudani**[2]

[1]Department of Computer Science, University of Manchester, Manchester, UK
[2]Human-Robot Interfaces and Interaction Laboratory, Italian Institute of Technology, Genoa, Italy

Abstract

Long-horizon humanoid loco-manipulation requires composing versatile whole-body skills and reliable high-level decision making. Existing methods often coordinate pretrained skills with scripted planners, finite-state machines or task-specific model-free policies, restricting their ability to handle complex task sequences. To address this limitation, we propose **LUCID**, a hierarchical model-based reinforcement learning framework that plans over reusable skills through imagined rollouts of a learned dynamics model. LUCID first trains a structured latent-conditioned low-level policy via adversarial imitation and then freezes it while jointly learning a high-level policy and macro-dynamics world model. The world model predicts the temporally extended state transitions induced by latent decisions, enabling high-level policy optimization through imagined rollouts. We evaluate our framework across various simulated multi-object rearrangement scenarios. Experimental results show that LUCID improves the full-task success and partial-completion rates compared to prior baseline methods, demonstrating its effectiveness in complex sequential loco-manipulation tasks.

Introduction

Humanoids can perform individual locomotion and manipulation behaviors, but executing them reliably in a long-horizon sequence remains a challenge (Gu et al. 2026). In multi-object rearrangement, for example, a humanoid needs to repeatedly navigate between objects and manipulate them within a continuous episode. Such tasks require both versatile whole-body skills and a high-level policy that can coordinate them toward sequential goals. Previous works have made substantial progress in physics-based humanoid control through adversarial imitation and latent skill learning, which enable natural whole-body motion and reusable motor behaviors (Peng et al. 2021, 2022; Weng et al. 2025; Wang et al. 2025a). Recent humanoid-scene interaction methods further extend such motor control to navigation and object manipulation in more complex environments (Starke et al. 2019; Hassan et al. 2023; Pan et al. 2024; Xiao et al. 2024). Together, these works have expanded humanoid control from isolated motor skills to scene-aware loco-manipulation.

*Corresponding author.

However, coordinating these behaviors over a long task sequence remains difficult. Existing approaches either consider primarily specialized controllers for isolated object interactions (Xu et al. 2024; Wang et al. 2025b), or connect consecutive interactions through hand-designed mechanisms, such as a fixed navigation-to-manipulation state machine (Pan et al. 2025) or per-object reference tracking with scripted handoffs between stages (Xu et al. 2025). Although such systems handle individual interactions effectively, their temporal organization is largely predefined and does not account for how intermediate decisions affect later subtasks. Hierarchical reinforcement learning addresses this structural problem by separating fast motor control from slower decision-making over temporally extended skills (Sutton, Precup, and Singh 1999). In physics-based character control, latent-skill methods have produced controllable motion representations and compact categorical priors that can be reused for downstream tasks (Tessler et al. 2023; Zhu et al. 2023). High-level policies can also combine pretrained humanoid motor primitives with limited task-specific reward design (Kuang et al. 2025). Yet existing hierarchical systems still rely on predefined skill transitions or task-specific model-free high-level policies trained through direct environment interaction. Consequently, they can select reusable behaviors but cannot predict how one skill changes the conditions for subsequent interactions, which is a limitation that becomes increasingly important in long-horizon multi-object rearrangement.

World models offer this predictive capability by approximating environment dynamics and using imagined trajectories to improve decision-making (Ha and Schmidhuber 2018; Hafner et al. 2020). They have demonstrated strong performance across control domains and have recently been applied to robotic manipulation, navigation, and locomotion (Wu et al. 2023; Li, Krause, and Hutter 2025). For humanoid interaction, however, predicting every joint-level transition over an extended horizon is difficult: high-dimensional, contact-rich dynamics require many autoregressive steps, amplifying model errors. The hierarchy provides a natural temporal abstraction. Rather than reproducing complete physical trajectories, a model can predict the task-level changes induced by each temporally extended skill. This connects reusable motor behaviors with predictive high-level reasoning, motivating a framework that learns skill-level dynamics and uses them to coordinate sequential humanoid interactions.

To this end, we introduce **LUCID**, a hierarchical model-based reinforcement learning framework for long-horizon skill composition, illustrated in Figure 1. Unlike hierarchical humanoid controllers that rely on scripted transitions or task-specific model-free skill sequencing, LUCID learns skill-level dynamics and optimizes a goal-conditioned high-level policy through imagined macro-transitions. A frozen latent-conditioned low-level controller provides reusable whole-body skills through a structured interface combining discrete skill anchors with continuous variation. At the macro timescale, the high-level policy selects latent commands, while a macro-dynamics world model predicts their task-level consequences for the humanoid, manipulated objects, and task progress rather than modeling every joint-level transition. Imagined rollouts through this model allow the policy to anticipate how current skill choices affect subsequent subtasks, while real simulator experience continually improves the model. This combination enables LUCID to coordinate complete multi-object rearrangement sequences without scripted handoffs.

In summary, our contributions are threefold.

- We propose **LUCID**, a hierarchical model-based reinforcement learning framework for task planning from imagined dynamics. Its world model predicts how latent decisions shape task progression, allowing a high-level policy to optimize long-horizon behavior in imagination.
- We train a low-level controller with adversarial imitation and an interaction-aware curriculum, using a structured latent interface that exposes whole-body motor primitives to the high-level policy.
- We evaluate our method in various long-horizon multi-object rearrangement tasks, showing improved full sequence completion and validating the key component through ablations.

Related Work

Physics-Based Motion Generation

Physics-based motion generation produces dynamically feasible behaviors by training controllers in simulation. DeepMimic popularized reference-tracking reinforcement learning for character animation (Peng et al. 2018), while AMP acquired task-compatible motion priors from unstructured demonstrations (Peng et al. 2021). Building on these foundations, reusable and directable representations enabled broader downstream applications (Peng et al. 2022; Tessler et al. 2023, 2024). Interaction-oriented research further expanded the scope to physical scene contacts, dynamic imitation, and object grasping (Hassan et al. 2023; Wang et al. 2023; Luo et al. 2024); general rearrangement (Xu et al. 2024); and scalable multi-skill execution (Xu et al. 2025; Pan et al. 2025). Recent systems have also transferred such capabilities to real humanoids (Wang et al. 2025a; Li et al. 2026). Unlike these methods, LUCID organizes phase-controllable skills with a high-level controller trained in imagination.

Hierarchical Reinforcement Learning

Hierarchical reinforcement learning decomposes long-horizon control into high-level decisions and low-level skills (Sutton, Precup, and Singh 1999; Bacon, Harb, and Precup 2017). Early systems built skill libraries and used schedulers to select or compose controllers (Faloutsos, Van de Panne, and Terzopoulos 2001; Heess et al. 2016). Later methods learned latent skill spaces that high-level policies could steer to generate motions (Merel et al. 2018; Peng et al. 2019; Hasenclever et al. 2020; Tessler et al. 2023). These abstractions were extended to legged loco-manipulation through skill sequencing and residual composition (Ji et al. 2022; Kumar, Essa, and Ha 2023). Recent humanoid systems instead plan motion references with hierarchical world models or blend goal-conditioned skills (Hansen et al. 2025; Kuang et al. 2025). LUCID learns discrete phase decisions over a frozen structured interface and optimizes long-horizon skill sequencing through imagined macro-dynamics for multi-object rearrangement.

World Model for Robotics

World models learn action-conditioned dynamics for planning and policy optimization through imagined trajectories (Ha and Schmidhuber 2018). PlaNet introduced latent dynamics for planning (Hafner et al. 2019), while Dreamer trained actor–critic policies from latent imagination (Hafner et al. 2020, 2025). TD-MPC combined learned representations with trajectory optimization (Hansen, Wang, and Su 2022; Hansen, Su, and Wang 2024), and DayDreamer extended imagination-based learning to physical robots (Wu et al. 2023). More recent robotics-specific models emphasize predictive reliability. RWM uses dual-autoregressive training to reduce compounding error and optimize locomotion policies in a learned neural simulator (Li, Krause, and Hutter 2025). HAIC instead predicts unobserved object dynamics from proprioception to improve feedback control during humanoid–object interaction (Li et al. 2026). However, these approaches primarily model joint-level dynamics, provide state estimates, or address continuous control. LUCID instead models skill-level transitions induced by structured phase commands and trains a high-level policy through imagined rollouts, enabling long-horizon planning while keeping the physics-based low-level controller fixed.

Method

Figure 1 presents LUCID, a scalable hierarchical framework for long-horizon robot-object interaction tasks. Training proceeds in two stages. First, we train a reusable latent-conditioned low-level controller (LLC) via adversarial imitation following ASE (Peng et al. 2022). We then freeze the LLC and jointly train a macro-dynamics world model (WM) and a goal-conditioned high-level controller (HLC) similar to DreamerV3 (Hafner et al. 2025). The WM predicts the physical consequences of latent decisions, allowing the HLC to learn from imagined macro-rollouts while real simulator interaction continually supplies training data for the WM.

Problem Formulation

We consider a simulated humanoid that must rearrange multiple objects in a prescribed order. We formulate the task as a goal-conditioned Markov Decision Process (MDP)

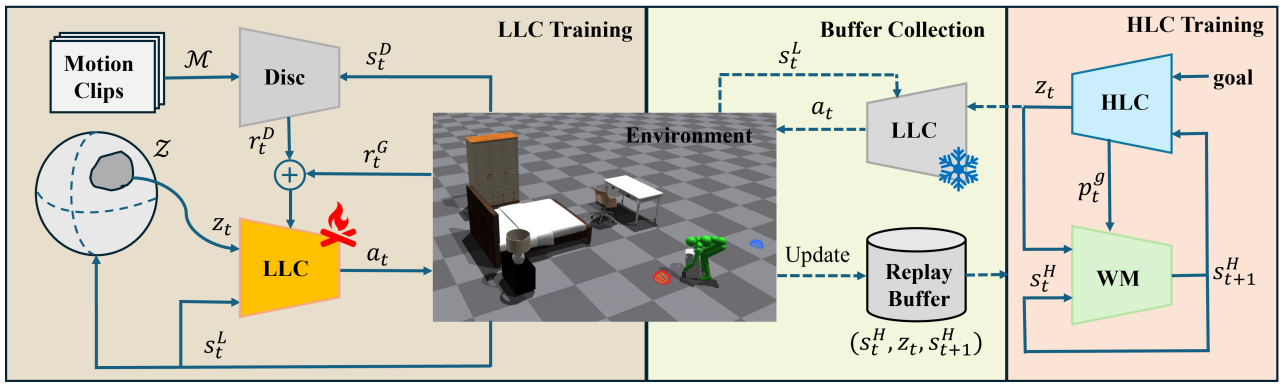


Figure 1: Overview of LUCID. LLC is trained with adversarial imitation and task rewards, then frozen. During training, latent z is produced with a state-based oracle. HLC is jointly trained with the macro-dynamics world model. The replay buffer is initialized with real transitions and refreshed during each update. Dashed lines indicate detached (stop-gradient) data flow.

$\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma, \mathcal{G} \rangle$. At simulation step t, $s_t \in \mathcal{S}$ contains the humanoid and environment states, $a_t \in \mathcal{A}$ denotes joint-level PD targets, and $\mathcal{T}$ is induced by the physics simulator. The episode goal $g = (g^1, \ldots, g^M) \in \mathcal{G}$ specifies an ordered sequence of M object-rearrangement subgoals. The objective is to maximize $\mathbb{E}[\sum_{t=0}^{T-1} \gamma^t r_t]$, where r_t rewards subgoal completion and physically plausible motion.

Latent Skill Primitives

The LLC $\pi_L(a_t \mid s_t^L, z_t)$ maps a low-level state s_t^L and a skill latent z_t to joint-level actions. The low-level state contains proprioception, object state, and the current task-guidance variables; in particular, the bounded guidance offset and waypoint-advance information represented by p_t^g are encoded into s_t^L before LLC execution. We train it to produce human-like, interaction-aware behaviors while remaining controllable through z_t by combining an adversarial imitation reward r_t^D with a goal-conditioned task reward r_t^G:

$$\max_{\pi_L} \mathbb{E}_{\tau \sim p(\tau \mid \pi_L, \psi)} \left[\sum_{t=0}^{T-1} \gamma^t \left(r_t^D + \lambda_G r_t^G \right) \right]. \tag{1}$$

Here, $\tau = (s_0^L, z_0, a_0, s_1^L, \ldots)$ is a low-level trajectory, λ_G balances task progress and imitation, and the oracle $\psi(s_t^L, g)$ selects the skill associated with the current interaction phase. By consistently conditioning each phase on its corresponding latent during training, the LLC learns to produce distinct behaviors for different latent inputs, thereby providing latent-level behavior controllability.

Adversarial imitation Let $\phi_t = \phi(s_t^D, s_{t+1}^D)$ be a transition feature constructed from robot and object kinematics. A discriminator $D(\phi) \in (0, 1)$ is trained to classify reference transitions as real and LLC transitions as generated. The LLC therefore receives the non-saturating reward:

$$r_t^D = -\log\big(1 - D(\phi_t)\big), \tag{2}$$

which increases when a generated transition is assigned a higher reference-motion probability. The full discriminator objective, including its gradient regularizer, is provided in the supplementary material (*Adversarial Imitation Objective*).

Structured latent space Unlike the unstructured hyperspherical prior used by ASE, we reserve one anchor for each of N semantic skills and use the remaining dimensions for within-skill variation:

$$z_t = \frac{[\,\mathbf{e}_{c_t};\ \sigma_z \boldsymbol{\epsilon}_t\,]}{\|[\,\mathbf{e}_{c_t};\ \sigma_z \boldsymbol{\epsilon}_t\,]\|_2}, \quad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{D_z - N}), \tag{3}$$

where $D_z \geq N$ is the latent dimension, $\mathbf{e}_{c_t} \in \mathbb{R}^N$ is the one-hot anchor for skill $c_t \in \{1, \ldots, N\}$, and σ_z controls within-skill variation. During LLC training, c_t is supplied by ψ, the induced latent distribution is thus a mixture $p(z_t) = \sum_{c=1}^{N} p(c_t = c)\, p(z_t \mid c_t = c)$ of N such neighborhoods rather than a uniform distribution over the sphere. After the LLC is frozen, the HLC selects the structured latent.

Interaction-aware curriculum To maintain behavior controllability during sequential object rearrangement, we train π_L using a three-stage curriculum: 1) *carry*, which emphasizes approach, contact, and lifting; 2) *rearrangement*, which enables transport and placement; and 3) *retreat and chaining*, which adds the transit skill, a subtask-completion bonus, and a post-placement retreat term. Each stage warm-starts from the previous one to target progressive completions, with only the task reward r^G and reference-state initialization (RSI) changing. At every stage, the task reward is defined as:

$$r_{t,m}^G = \sum_j w_j^{(m)} r_t^{(j)}, \tag{4}$$

where m indexes the curriculum stage, $r_t^{(j)}$ is an interaction reward, and $w_j^{(m)}$ is its stage-dependent coefficient. The stage-specific objective, retreat reward, activation conditions, and initialization scheme are given in the supplementary material (*Interaction-Aware Curriculum*).

Macro-Dynamics World Model

Macro transition The WM predicts task-level physical consequences over K low-level control steps. At macro step t, the HLC chooses $a_t^H = (z_t, p_t^g)$, where z_t is the structured skill latent and p_t^g contains a bounded guidance offset and

a waypoint-advance gate. Before LLC execution, p_t^g is encoded into the task-guidance portion of s_t^L. The frozen LLC then executes the selected latent for up to K simulator steps, producing one real macro transition. We decompose the compact macro state as $s_t^H = (s_t^c, s_t^f)$, where s_t^c contains continuous task variables (e.g., poses, velocities, displacements, lift heights) and s_t^f contains binary task-progress flags. The dynamics network uses (s_t^H, a_t^H) to predict the next continuous state and progress flags, whereas a separate continuation head estimates whether the episode remains active:

$$\begin{aligned} \hat{s}_{t+1}^c &= s_t^c + f_\theta(s_t^H, a_t^H), \\ \hat{s}_{t+1}^f &= \mathrm{sigmoid}\big(g_\theta(s_t^H, a_t^H)\big), \\ \hat{\zeta}_t &= \mathrm{sigmoid}\big(h_\theta(s_t^H)\big). \end{aligned} \tag{5}$$

Here, f_θ predicts the change in the continuous variables rather than their absolute next values. This residual formulation focuses learning on the state change induced by one macro action. In contrast, g_θ directly predicts the probabilities of the binary flags because their abrupt changes, such as a placed flag switching from zero to one, are poorly represented by continuous residuals. The continuation prediction $\hat{\zeta}_t \in [0, 1]$ estimates the probability that the transition does not terminate because of failure; it subsequently discounts value propagation through imagined trajectories.

Training process First, we pretrain the world model with real macro-transitions collected from the simulator in a supervised fashion. We prefill a replay buffer with real macro transitions generated by an exploratory policy π_{rand} that samples structured latents and guidance variables, and run a fixed number of single-step regression updates. The world model is optimized with the following loss:

$$\begin{aligned} \mathcal{L}_{\mathrm{WM}} = \frac{1}{d_c} \left\| \hat{s}_{t+1}^c - s_{t+1}^c \right\|_2^2 \\ + \lambda_f \,\mathrm{BCE}\left(\hat{s}_{t+1}^f, s_{t+1}^f\right) + \lambda_\zeta \,\mathrm{BCE}\left(\hat{\zeta}_t, \zeta_t\right). \end{aligned} \tag{6}$$

Here, the first term is the mean-squared error on the continuous part, the second term is a binary cross-entropy on the discrete task flags, and the last term trains the continuation head. Afterwards, we train the world model and π_H concurrently using a shared replay buffer. Each iteration includes three steps: (i) rolling out the current policy in the simulator to collect macro-transitions and adding them to the buffer; (ii) updating the world model on mini-batches sampled from the buffer; and (iii) updating π_H on short rollouts imagined by the newly updated world model. Note that the buffer is constantly refreshed with the real experience visited by the current π_H at each update. The supplementary material (*Macro-Dynamics World Model*) specifies the complete loss, target construction, replay-buffer procedure, and rollout convention.

High-Level Learning in Imagined Dynamics

The goal-conditioned actor $\pi_{H,\varphi}(a_t^H \mid s_t^H, g)$ and critic $V_\xi(s_t^H, g)$ operate at the macro timescale. At each update, real start states are sampled from $\mathcal{B}$ and the actor and WM are unrolled for Q imagined macro steps. The high-level reward $r_t^H = r^H(s_t^H, a_t^H, \hat{s}_{t+1}^H, g)$ uses the same task-progress definition for real and imagined transitions. Let $v_t = V_\xi(s_t^H, g)$. We estimate imagined returns using:

$$V_t^\lambda = r_t^H + \gamma \hat{\zeta}_t \big[(1 - \lambda_{\mathrm{ret}}) v_{t+1} + \lambda_{\mathrm{ret}} V_{t+1}^\lambda\big], \quad V_Q^\lambda = v_Q, \tag{7}$$

where λ_{ret} is the return-trace parameter and $\hat{\zeta}_t$ downweights returns across predicted termination. Because the high-level action includes sampled discrete decisions, we optimize the actor with a score-function estimator:

$$\mathcal{L}_{\mathrm{actor}} = \mathbb{E}_{\hat{\tau}} \left[\sum_{t=0}^{Q-1} d_t \ell_t^{\mathrm{actor}} \right], \tag{8}$$

$$\ell_t^{\mathrm{actor}} = -\,\mathrm{sg}(A_t) \log \pi_{H,\varphi}(a_t^H \mid s_t^H, g) - \eta_H \mathcal{H}_t, \tag{9}$$

where $A_t = (V_t^\lambda - v_t)/\max(S, \varepsilon)$ is a return-normalized advantage, $d_t = \prod_{k=0}^{t-1} \gamma \hat{\zeta}_k$ is the cumulative continuation discount, S is a running return scale, $\varepsilon > 0$ prevents division by zero, $\mathcal{H}_t$ is the entropy of the sampled high-level action distribution, and sg denotes stop-gradient. The critic is trained toward detached V_t^λ targets with a distributional value loss, while an exponential-moving-average target critic provides a detached distributional regularizer. To stabilize early skill discovery, we warm-start the categorical skill-selection component from a geometry-based oracle. Its behavior-cloning coefficient is annealed to zero, so the oracle initializes rather than fixes the final controller.

Algorithm 1 summarizes the full optimization procedure for LUCID. Here, $\mathcal{J}_L$ is the LLC objective in Eq. (1), $\mathcal{M}_\theta$ is the macro-dynamics world model, and θ, φ, and ξ denote the WM, actor, and critic parameters, respectively. The

Algorithm 1: LUCID training pipeline

Require: Motion data $\mathcal{M}$, simulator Ψ, oracle ψ, macro period K
Ensure: Frozen LLC π_L and trained HLC $\pi_{H,\varphi}$

$\pi_L \leftarrow \arg\max_\pi \mathcal{J}_L(\pi; \mathcal{M}, \Psi, \psi)$; freeze π_L
Initialize WM $\mathcal{M}_\theta$, actor $\pi_{H,\varphi}$, critic V_ξ, target critic $V_{\bar{\xi}}$, and replay buffer $\mathcal{B}$
for $i = 1, \ldots, N_{\mathrm{pre}}$ **do**
  $\tau^{\mathrm{real}} \leftarrow \mathsf{Collect}_K(\pi_L, \pi_{\mathrm{rand}})$
  Label visited states with ψ and insert τ^{real} into $\mathcal{B}$
end for
for $i = 1, \ldots, N_{\mathrm{WM}}$ **do**
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\mathrm{WM}}(\theta; \mathsf{Sample}(\mathcal{B}))$
end for
repeat
  $\tau^{\mathrm{real}} \leftarrow \mathsf{Collect}_K(\pi_L, \pi_{H,\varphi})$
  Label its visited states with ψ and refresh $\mathcal{B}$
  $\theta \leftarrow \theta - \eta_\theta \nabla_\theta \mathcal{L}_{\mathrm{WM}}(\theta; \mathsf{Sample}(\mathcal{B}))$
  $\mathcal{D} \sim \mathcal{B}$
  $\hat{\tau} \leftarrow \mathsf{Imagine}_Q(\mathcal{D}, \pi_{H,\varphi}, \mathcal{M}_\theta)$
  $\xi \leftarrow \xi - \eta_\xi \nabla_\xi \mathcal{L}_{\mathrm{critic}}(\xi; \hat{\tau})$
  $\varphi \leftarrow \varphi - \eta_\varphi \nabla_\varphi \left[\mathcal{L}_{\mathrm{actor}}(\varphi; \hat{\tau}) + \beta_{\mathrm{BC},k} \mathcal{L}_{\mathrm{BC}}(\varphi; \mathcal{D}, \psi)\right]$
  $\bar{\xi} \leftarrow (1 - \alpha)\bar{\xi} + \alpha\xi$; $\beta_{\mathrm{BC},k+1} \leftarrow \mathsf{Decay}(\beta_{\mathrm{BC},k})$
until the training budget is exhausted
return $\pi_L, \pi_{H,\varphi}$

oracle-labeled buffer is used only while $\beta_{\mathrm{BC},k} > 0$. The exact joint action distribution, critic loss, oracle objective, and actor–WM update schedule are provided in the supplementary material (*High-Level Controller Training*).

Experiments

We evaluate LUCID on long-horizon multi-object rearrangement, focusing on three questions: whether it improves sequential task completion over representative baselines, whether it remains effective on held-out layouts and longer task chains, and which components account for its performance. We first define the shared evaluation protocol, then report comparison results for diverse settings, followed by world-model and latent-interface ablations.

Experimental Setup

Dataset We construct multi-object rearrangement tasks from HITR (Xu et al. 2024). Movable objects are initialized on the ground in diverse physically feasible configurations across warehouse, kitchen, bedroom, and living-room layouts. We compose single-object instances into sequential episodes in which each object must be placed before the next subtask becomes active. The ID split contains 62 training tasks, whereas the OOD split contains 20 tasks. The standard benchmark contains two-object chains; a separate extension evaluates chains of up to five objects. LLC training additionally uses reference motions from OMOMO (Li, Wu, and Liu 2023) and SAMP (Hassan et al. 2021).

Metrics We report three metrics. **Success Rate** SR_k is the percentage of episodes in which the first k objects are placed within $0.2\,\mathrm{m}$ of their targets. **Average Placement Error** (APE) is the final object-to-goal distance averaged over all task objects. **Time** is the mean simulated completion time over successful episodes. Values are mean $\pm$ sample standard deviation over three evaluation seeds using the same task split and success criterion for every method.

Baselines We compare against the following baselines using task-specific deployment adapters. **InterMimic** (Wang et al. 2025b) is a motion-tracking policy; for each subtask, we construct a kinematic human–object reference by warping motion segments to the required waypoints and object poses. **TokenHSI** (Pan et al. 2025) uses a scripted finite-state adapter to switch between locomotion and object-carrying skills. **HumanVLA** (Xu et al. 2024) is an end-to-end single-object rearrangement policy, which we extend by activating subtasks sequentially. All methods use the same ID/OOD task lists, simulator horizon, placement threshold, and evaluation seeds. Because these baselines do not model autonomous inter-subtask transitions, their adapters reset only the humanoid to a favorable pose at each handoff. Consequently, SR and APE compare LUCID against favorably initialized baseline deployments, while completion times are descriptive and not directly comparable.

Implementation Details

We use Isaac Gym (Makoviychuk et al. 2021) and a humanoid with 15 rigid bodies and 28 PD-controlled joints (Hassan et al. 2023). The LLC follows ASE but replaces its learned skill encoder and diversity objective with the structured latent interface. We train it with PPO (Schulman et al. 2017) and a staged curriculum that progresses from robust single-object interaction to two-object chaining. The HLC uses a DreamerV3-style actor–critic, and the deterministic world model is a plain MLP over the compact task state. The simulator runs at $60\,\mathrm{Hz}$, the LLC acts at $30\,\mathrm{Hz}$, and the HLC selects one macro action every 20 LLC steps. The HLC uses a 12-step imagination horizon, sequence length 32, and batch size 64. We train the LLC on two NVIDIA RTX 4090 GPUs and train the HLC and world model jointly with 8192 parallel environments. Further architecture, optimization, and evaluation details are provided in the supplementary material.

Comparison Experiments

Quantitative evaluation Table 1 compares all methods on the shared ID and OOD protocols. LUCID obtains the highest mean SR_1, SR_2, and the lowest APE on both splits despite receiving no handoff reset. On ID tasks, its SR_2 exceeds the strongest baseline by 33.6 percentage points (73.4% versus 39.8%); on OOD tasks, the margin is 31.4 points (68.4% versus 37.0%). The smaller SR_1–SR_2 gap for LUCID (15.8 points ID and 15.5 points OOD) indicates more reliable handoffs than the baselines. From ID to OOD, LUCID decreases by 5.3 points in SR_1 and 5.0 points in SR_2, while retaining the best mean placement accuracy. InterMimic performs worst overall, consistent with the difficulty of constructing precise warped references for diverse object configurations. Baseline completion times are shorter because their handoff adapters teleport the humanoid; we therefore treat Time as descriptive rather than a direct efficiency comparison.

Qualitative performance Figure 2 shows LUCID completing multi-object rearrangement chains in diverse layouts. We can see that LUCID exhibits clean phase structures and well-separated transitions across all scenes, with HLC producing interpretable skills that are faithfully executed by the frozen LLC. The same latent repertoire generalizes across layouts without per-scene tuning, adapting to environmental obstacles while completing the task chain fully autonomously. These are consistent with the SR2 and placement-accuracy gains reported in Table 1. The supplementary material provides the corresponding evaluation protocol and implementation details.

Extended multi-objects tasks To assess long-horizon robustness beyond the two-object training horizon, we evaluate chains of up to five objects and report prefix success SR_k in Figure 3. InterMimic and TokenHSI approach zero by SR_4, and HumanVLA retains 5% success at SR_5. This rapid decay reflects their reliance on scripted planners, which accumulate handoff errors without recovery. However, LUCID degrades more gracefully, achieving around 56% at SR3 and still 21% at SR5. We attribute this to the joint evolution of the world model and high-level policy during training, which together yields a more reliable long-horizon planner.

Method	In-Distribution				Out-of-Distribution			
	SR_1 (%) ↑	SR_2 (%) ↑	APE (m) ↓	Time (s) ↓	SR_1 (%) ↑	SR_2 (%) ↑	APE (m) ↓	Time (s) ↓
HumanVLA	79.4 ± 0.7	39.8 ± 0.8	0.83 ± 0.02	6.1 ± 0.1	77.8 ± 5.6	37.0 ± 0.8	0.77 ± 0.05	6.2 ± 0.4
InterMimic	23.7 ± 4.6	4.7 ± 1.2	2.21 ± 0.04	10.4 ± 1.3	19.4 ± 7.3	3.7 ± 3.2	2.11 ± 0.23	11.9 ± 1.9
TokenHSI	69.4 ± 1.1	28.7 ± 1.9	0.96 ± 0.15	7.6 ± 0.3	66.1 ± 3.7	26.4 ± 3.5	1.17 ± 0.09	8.2 ± 0.2
LUCID	89.2 ± 0.7	73.4 ± 3.0	0.32 ± 0.02	15.4 ± 0.9	83.9 ± 0.8	68.4 ± 3.8	0.52 ± 0.05	18.3 ± 2.4

Table 1: Comparison on the shared ID and OOD task splits. SR_k denotes completion through subtask k.

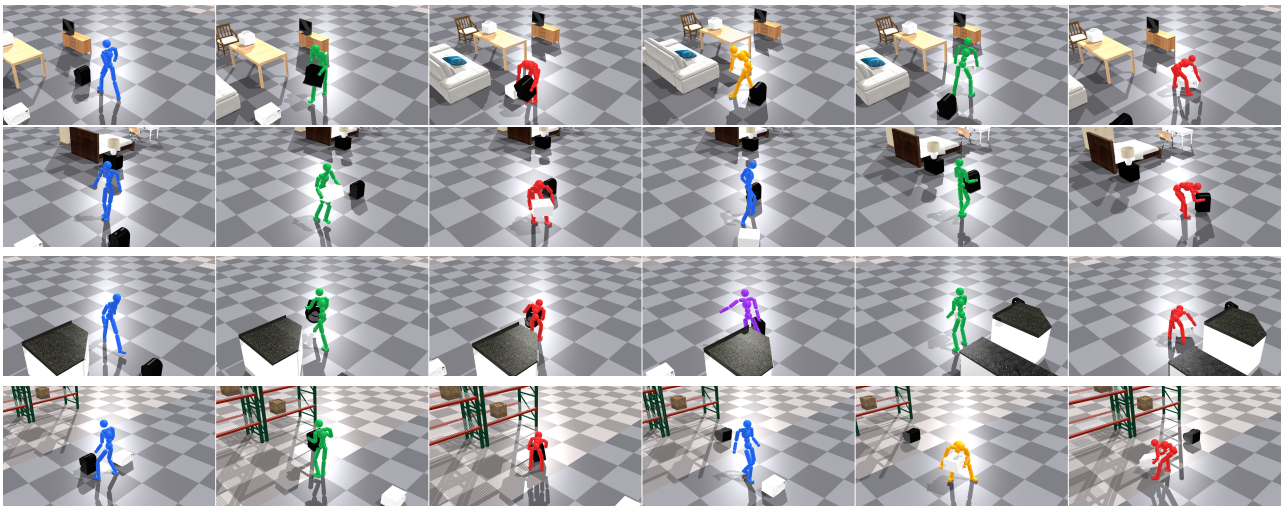


Figure 2: Qualitative filmstrips of LUCID in various task layouts (top to bottom: living room, bedroom, kitchen, warehouse). Colors indicate semantic latent decisions: locomotion (blue), transport (green), release (red), fetch (yellow), and transit (purple). The frames illustrate continuous navigation, object interaction, and autonomous handoff without resetting the humanoid.

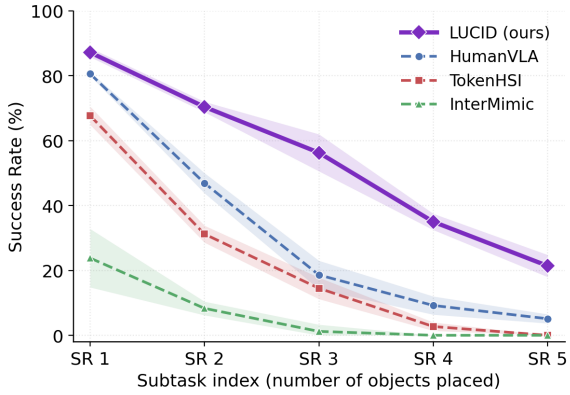


Figure 3: Success rates against the number of objects.

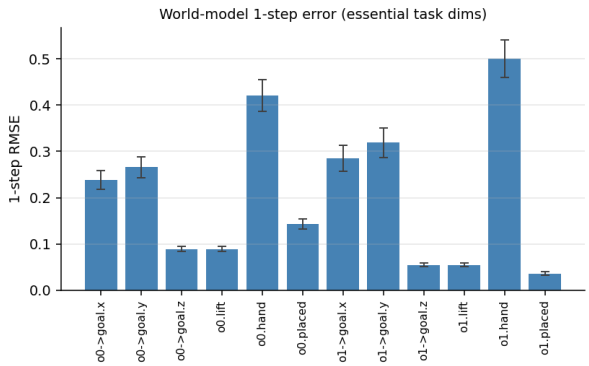


Figure 4: World-model one-step RMSE on task-relevant compact-state dimensions. Bars average per-environment RMSE and whiskers denote the standard error across evaluation environments.

Ablation Studies

We investigate the key components of LUCID by addressing the following questions: 1) *Does the world model improve HLC learning through imagined rollouts?* 2) *Does the structured latent interface improve behavior controllability?*

On world model effectiveness We first evaluate predictive accuracy independently of downstream task success. Figure 4 reports one-step errors for task-critical dimensions, including object–goal displacement, lift height, hand–object distance, and placement progress. Figure 5 then visualizes a representative 12-step open-loop rollout. The supplementary material specifies the action sequence, sampling protocol, and aggregate horizon-wise errors used alongside this qualitative trajectory. Table 2 compares HLCs trained with world-model

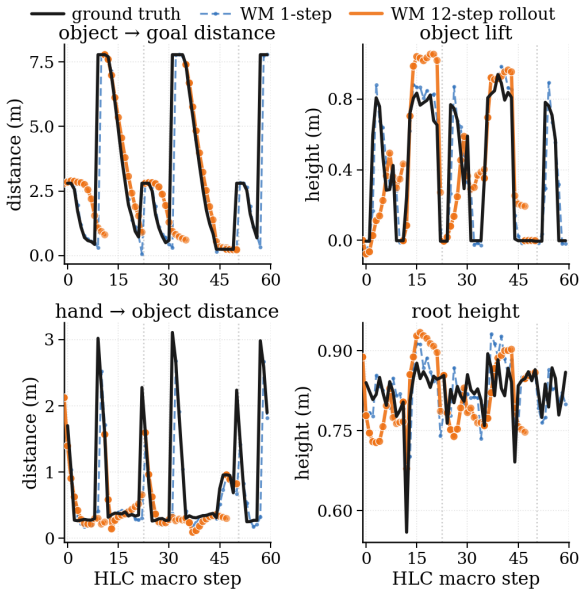


Figure 5: Multi-step world-model predictions. Black curves show simulator states, dashed blue curves show one-step predictions, and orange curves show imagined rollouts.

Method	ID		OOD	
	SR1 (%) ↑	SR2 (%) ↑	SR1 (%) ↑	SR2 (%) ↑
MF-HLC-sparse	84.8 ± 0.7	59.3 ± 1.3	82.6 ± 0.8	42.8 ± 1.1
MF-HLC-dense	87.8 ± 0.6	61.9 ± 1.1	65.0 ± 2.0	33.1 ± 1.2
LUCID (*sparse*)	89.5 ± 1.0	73.5 ± 3.0	81.9 ± 0.9	63.1 ± 2.7
LUCID-dense	90.8 ± 0.7	74.0 ± 1.0	70.1 ± 0.9	50.2 ± 3.2

Table 2: WM ablation under sparse and dense rewards.

(WM) imagination against model-free (MF) learning under both sparse and dense reward settings. The WM variants achieve higher mean SR_2 in all four split–reward comparisons: 73.5 versus 59.3 and 74.0 versus 61.9 on ID, and 63.1 versus 42.8 and 50.2 versus 33.1 on OOD. This consistent advantage supports the use of imagined macro-transitions for learning long-horizon task progression. Dense shaping yields only a 0.5-point ID SR_2 gain for LUCID but reduces OOD SR_2 by 12.9 points, suggesting weaker transfer than the sparse setting.

On structured latent interface Figure 6 separates command decodability from task relevance. We drive each frozen LLC with five commands in the same two-object environment, retain states in which the object is initially reachable, and describe the resulting motions with identical kinematic features. The independently fitted t-SNE plots and five-fold random-forest probe show that both interfaces produce command-decodable behaviors (balanced accuracy 0.86 and 0.95). Decodability alone is therefore insufficient. The manipulation panel instead shows that three structured commands exceed the sustained-lift threshold, whereas none of the unstructured commands does. Table 3 further shows that the structured interface reaches 74.2% ID and 66.4% OOD SR_2, while the unstructured interface obtains 0.0% SR_2 on both splits. These results indicate that the relevant property is not merely diverse behavior, but a command set aligned with the interaction stages required by the downstream task.

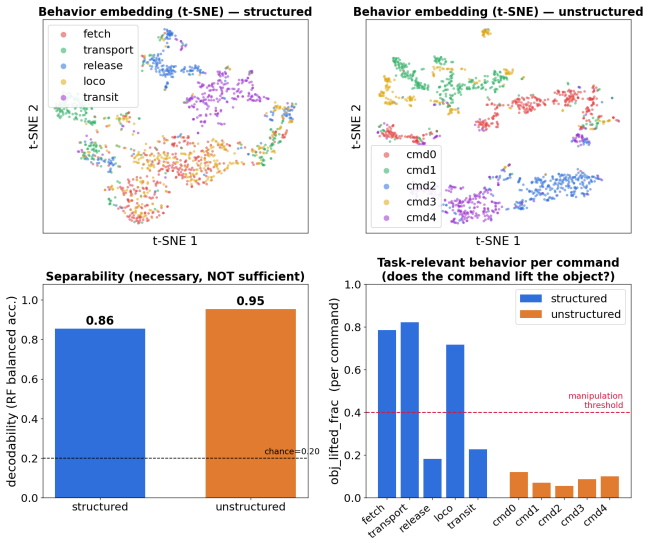


Figure 6: Behavior controllability of the structured and unstructured latent interface. Top: separately fitted t-SNE visualizations of identical kinematic descriptors. Bottom left: five-fold random-forest balanced accuracy for command decoding. Bottom right: mean fraction of time each command lifts the object, with 0.4 used as the sustained threshold.

Method	ID		OOD	
	SR1 (%) ↑	SR2 (%) ↑	SR1 (%) ↑	SR2 (%) ↑
LUCID (*struct*)	87.5 ± 1.5	74.2 ± 2.6	82.0 ± 1.5	66.4 ± 3.6
LUCID-unstruct	4.8 ± 2.6	0.0 ± 0.0	2.7 ± 1.9	0.0 ± 0.0

Table 3: Structured-interface ablation. Each HLC is trained and evaluated with its corresponding frozen LLC interface under the same ID/OOD protocol.

Conclusion

In this paper, we introduced LUCID, a hierarchical framework for multi-object rearrangement with a simulated humanoid. It coordinates a structured low-level controller using a high-level policy trained through imagined skill-level transitions, improving long-horizon completion. Results show that task-aligned temporal abstraction is central: a structured skill interface makes high-level decisions meaningful, while macro-level prediction supports long-horizon reasoning without reproducing joint-level dynamics. Current limitations include privileged state inputs and limited task, object, and embodiment diversity; future work will address visual control, broader tasks, and real-world transfer.

References

Bacon, P.-L.; Harb, J.; and Precup, D. 2017. The option-critic architecture. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Faloutsos, P.; Van de Panne, M.; and Terzopoulos, D. 2001. Composable controllers for physics-based character animation. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 251–260.

Gu, Z.; Li, J.; Shen, W.; Yu, W.; Xie, Z.; McCrory, S.; Cheng, X.; Shamsah, A.; Griffin, R.; Liu, C. K.; et al. 2026. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. *IEEE/ASME Transactions on Mechatronics*, 31(2): 2300–2330.

Ha, D.; and Schmidhuber, J. 2018. World models. *arXiv preprint arXiv:1803.10122*, 2(3): 440.

Hafner, D.; Lillicrap, T.; Ba, J.; and Norouzi, M. 2019. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*.

Hafner, D.; Lillicrap, T.; Norouzi, M.; and Ba, J. 2020. Mastering atari with discrete world models.

Hafner, D.; Pasukonis, J.; Ba, J.; and Lillicrap, T. 2025. Mastering diverse control tasks through world models. *Nature*, 640(8059): 647–653.

Hansen, N.; Su, H.; and Wang, X. 2024. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, 47376–47405.

Hansen, N.; SV, J.; Sobal, V.; LeCun, Y.; Wang, X.; and Su, H. 2025. Hierarchical world models as visual whole-body humanoid controllers. In *International Conference on Learning Representations*, volume 2025, 62175–62195.

Hansen, N.; Wang, X.; and Su, H. 2022. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*.

Hasenclever, L.; Pardo, F.; Hadsell, R.; Heess, N.; and Merel, J. 2020. CoMic: Complementary task learning & mimicry for reusable skills. In *International Conference on Machine Learning*, 4105–4115. PMLR.

Hassan, M.; Ceylan, D.; Villegas, R.; Saito, J.; Yang, J.; Zhou, Y.; and Black, M. J. 2021. Stochastic scene-aware motion prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11374–11384.

Hassan, M.; Guo, Y.; Wang, T.; Black, M.; Fidler, S.; and Peng, X. B. 2023. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*, 1–9.

Heess, N.; Wayne, G.; Tassa, Y.; Lillicrap, T.; Riedmiller, M.; and Silver, D. 2016. Learning and transfer of modulated locomotor controllers. *arXiv preprint arXiv:1610.05182*.

Ji, Y.; Li, Z.; Sun, Y.; Peng, X. B.; Levine, S.; Berseth, G.; and Sreenath, K. 2022. Hierarchical reinforcement learning for precise soccer shooting skills using a quadrupedal robot. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1479–1486. IEEE.

Kuang, Y.; Geng, H.; Elhafsi, A.; Do, T.-D.; Abbeel, P.; Malik, J.; Pavone, M.; and Wang, Y. 2025. Skillblender: Towards versatile humanoid whole-body loco-manipulation via skill blending. *arXiv preprint arXiv:2506.09366*.

Kumar, K. N.; Essa, I.; and Ha, S. 2023. Cascaded compositional residual learning for complex interactive behaviors. *IEEE Robotics and Automation Letters*, 8(8): 4601–4608.

Li, C.; Krause, A.; and Hutter, M. 2025. Robotic world model: A neural network simulator for robust policy optimization in robotics. *arXiv preprint arXiv:2501.10100*.

Li, D.; Chen, X.; Wu, Q.; Chen, B.; Wu, S.; Wu, H.; Zhang, G.; Li, L.; Zhou, M.; Xiang, D.; et al. 2026. Haic: Humanoid agile object interaction control via dynamics-aware world model. *arXiv preprint arXiv:2602.11758*.

Li, J.; Wu, J.; and Liu, C. K. 2023. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)*, 42(6): 1–11.

Luo, Z.; Cao, J.; Christen, S.; Winkler, A.; Kitani, K.; and Xu, W. 2024. Omnigrasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems*, 37: 2161–2184.

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

Merel, J.; Hasenclever, L.; Galashov, A.; Ahuja, A.; Pham, V.; Wayne, G.; Teh, Y. W.; and Heess, N. 2018. Neural probabilistic motor primitives for humanoid control. *arXiv preprint arXiv:1811.11711*.

Pan, L.; Wang, J.; Huang, B.; Zhang, J.; Wang, H.; Tang, X.; and Wang, Y. 2024. Synthesizing physically plausible human motions in 3d scenes. In *2024 International Conference on 3D Vision (3DV)*, 1498–1507. IEEE.

Pan, L.; Yang, Z.; Dou, Z.; Wang, W.; Huang, B.; Dai, B.; Komura, T.; and Wang, J. 2025. Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5379–5391.

Peng, X. B.; Abbeel, P.; Levine, S.; and Van de Panne, M. 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)*, 37(4): 1–14.

Peng, X. B.; Chang, M.; Zhang, G.; Abbeel, P.; and Levine, S. 2019. Mcp: Learning composable hierarchical control with multiplicative compositional policies. *Advances in neural information processing systems*, 32.

Peng, X. B.; Guo, Y.; Halper, L.; Levine, S.; and Fidler, S. 2022. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4): 1–17.

Peng, X. B.; Ma, Z.; Abbeel, P.; Levine, S.; and Kanazawa, A. 2021. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4): 1–20.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Starke, S.; Zhang, H.; Komura, T.; and Saito, J. 2019. Neural state machine for character-scene interactions. *ACM Transactions on Graphics*, 38(6): 178.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2): 181–211.

Tessler, C.; Guo, Y.; Nabati, O.; Chechik, G.; and Peng, X. B. 2024. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions On Graphics (TOG)*, 43(6): 1–21.

Tessler, C.; Kasten, Y.; Guo, Y.; Mannor, S.; Chechik, G.; and Peng, X. B. 2023. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 conference proceedings*, 1–9.

Wang, H.; Zhang, W.; Yu, R.; Huang, T.; Ren, J.; Jia, F.; Wang, Z.; Niu, X.; Chen, X.; Chen, J.; et al. 2025a. Physhsi: Towards a real-world generalizable and natural humanoid-scene interaction system. *arXiv preprint arXiv:2510.11072*.

Wang, Y.; Lin, J.; Zeng, A.; Luo, Z.; Zhang, J.; and Zhang, L. 2023. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393*.

Wang, Y.; Zhao, Q.; Yu, R.; Tsui, H. W.; Zeng, A.; Lin, J.; Luo, Z.; Yu, J.; Li, X.; Chen, Q.; et al. 2025b. Skillmimic: Learning basketball interaction skills from demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17540–17549.

Weng, H.; Li, Y.; Sobanbabu, N.; Wang, Z.; Luo, Z.; He, T.; Ramanan, D.; and Shi, G. 2025. Hdmi: Learning interactive humanoid whole-body control from human videos. *arXiv preprint arXiv:2509.16757*.

Wu, P.; Escontrela, A.; Hafner, D.; Abbeel, P.; and Goldberg, K. 2023. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, 2226–2240. PMLR.

Xiao, Z.; Wang, T.; Wang, J.; Cao, J.; Zhang, W.; Dai, B.; Lin, D.; and Pang, J. 2024. Unified human-scene interaction via prompted chain-of-contacts. In *International Conference on Learning Representations*, volume 2024, 24450–24461.

Xu, S.; Ling, H. Y.; Wang, Y.-X.; and Gui, L.-Y. 2025. Intermimic: Towards universal whole-body control for physics-based human-object interactions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12266–12277.

Xu, X.; Zhang, Y.; Li, Y.-L.; Han, L.; and Lu, C. 2024. Humanvla: Towards vision-language directed object rearrangement by physical humanoid. *Advances in Neural Information Processing Systems*, 37: 18633–18659.

Zhu, Q.; Zhang, H.; Lan, M.; and Han, L. 2023. Neural categorical priors for physics-based character control. *ACM Transactions on Graphics (TOG)*, 42(6): 1–16.