CO^{2}SKILL: WHOLE-BODY CONTROL VIA SKILL COMPOSITION FOR LONG-HORIZON HUMAN-ENVIRONMENT INTERACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Achieving human-level dexterity in complex, unstructured environments requires the seamless integration of whole-body scene interaction and dexterous object manipulation skills. While existing physics-based controllers generate physically plausible behaviors in each domain, they largely address these two capabilities independently. In this paper, we present Co^{2}Skill that integrates scene interaction and dexterous manipulation through a unified policy formulation. Built on a pretrained motion prior, the policy uses task and phase dependent observation masks to select information relevant to the current interaction goals. We introduce a goal-conditioned loco-manipulation curriculum that combines partial reference guidance for precision with exploration from varied initial states while allowing goal-directed execution beyond the demonstrated trajectories. We further introduce a cross-task curriculum that jointly trains individual skills and selected task sequences, preserving physical states across task boundaries and maintaining grasps during subsequent scene interactions. Together, these support sequential task execution and simultaneous scene interaction with object manipulation. We evaluate sitting, standing, climbing, stair traversal, and goal-directed manipulation, together with sequential execution and with random different conditions. Additionally, we demonstrate skill compositions in indoor environments, illustrating their integration within the same control formulation.

1 INTRODUCTION

Everyday activities at home rarely consist of a single skill. A person may walk to a shelf, pick up a phone, and then sit down on a chair while still holding it. For a simulated humanoid, this means coordinating locomotion, whole-body interaction with the surrounding scene, and finger-level grasping across a room-scale space rather than on an empty floor or a tabletop. Each skill must also remain physically stable when it starts from whatever state the previous skill leaves behind.

Physics-based character control with deep reinforcement learning has made steady progress toward such behavior. Tracking policies trained on large motion capture collections now imitate a wide range of human movement (Peng et al., 2018; Luo et al., 2023; Tessler et al., 2024), and motion priors distilled from these policies can be reused for new downstream tasks (Luo et al., 2024). Building on such learned motor skills, two lines of work have developed largely in isolation. Human-scene interaction (HSI) controllers move the whole body relative to furniture and terrain, for example to sit on a chair or climb a staircase (Xiao et al., 2024; Pan et al., 2025), while the hands (with no fingers attached) play little role. Human-object interaction (HOI) controllers achieve precise finger contact and object control (Luo et al., 2025; Tessler et al., 2025), but they typically operate close to the object and often depend on dense reference trajectories from datasets such as GRAB (Taheri et al., 2020). Ideally, a single controller should both interact with the surrounding scene and manipulate the objects within it.

Combining these capabilities is harder than simply putting the two kinds of skills together. First, scene interaction and manipulation depend on different inputs. Scene interaction needs scene geometry and spatial targets, while manipulation needs object motion, hand poses, and contact. Second, each skill usually learns from a fixed set of starting states, which differ from the states it meets in a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: Our method effectively deals with both scene interactions and dexterous manipulations.

longer task. A sitting policy that learns from a normal standing pose never starts with an object in its hands, and a grasping policy mostly starts from the pose in its reference clip. Switching between separate specialist policies does not solve this problem, and it cannot handle cases where a scene task and a grasp must be performed at the same time. Third, reference motions give the detailed guidance needed for precise contact or large body adjustment, but they cover only the situations present in the motion capture data. A controller that relies on them therefore struggles to approach a chair or an object from new positions.

We address these challenges with our Co^{2}Skill, a physics-based controller that performs scene interaction and dexterous manipulation with a single policy. We build the policy on top of a pretrained motion prior and decoder (Luo et al., 2024). To handle the different inputs of each task, we design a transformer that reads task tokens for the character, scene, object, and hands, enabling only those relevant to the current task and phase. To handle the gap between training and initial states and the limited coverage of reference motions, we introduce two curricula. The goal-conditioned loco-manipulation curriculum uses reference motions to guide close-range interactions such as sitting down or forming a grasp, while exploration from varied starting states teaches the policy to reach the same goals beyond the reference. The cross-task curriculum trains individual tasks together with task sequences, such as lifting an object and then sitting while holding it, so that the policy learns from the states that sequential execution actually produces. Our shared policy outperforms MaskedMimic across five scene-interaction tasks and achieves success rates matching or exceeding per-task specialists. We also evaluate pick-move-place task and qualitatively demonstrate manipulation-to-scene transitions while maintaining object grasps.

Our contributions are summarized as follows:

- **A Unified Framework for Scene Interaction and Dexterous Manipulation.** We introduce a unified control framework that combines whole-body scene interaction and dexterous object manipulation within a single goal-conditioned policy.
- **A Goal-Conditioned Loco-Manipulation Curriculum.** We combine reference-guided interaction learning with exploration from varied initial states, using partial motion guidance to learn precise interactions while allowing goal-directed execution beyond the reference trajectory.
- **A Cross-Task Curriculum for Sequential and Simultaneous Interactions.** We jointly train individual tasks and selected task sequences, preserving physical states across transitions and maintaining grasp conditions during subsequent scene interactions.

2 Related Work

2.1 Human-Scene Interactions

Understanding human-scene interactions requires datasets that faithfully capture spatial relations between human motion and environmental context. Early mocap datasets of humans interacting with static environments (Starke et al., 2019; Hassan et al., 2021) enabled training models to predict behavior given goal targets. Interaction Replica (Guzov et al., 2022) extended this by tracking interactions with partially dynamic scenes. Locomotion-Action-Manipulation (Lee & Joo, 2023) synthesizes human-scene interactions in 3D environments by coordinating locomotion with manipulation actions. More recent work augments datasets by capturing human behavior and applying VR rendering to diversify indoor environments (Jiang et al., 2024b;a). These datasets establish foundations for learning human behavior in environments.

2.2 Human-Object Interactions

Human behavior naturally incorporates manipulation of objects in the surrounding environment. BEHAVE (Bhatnagar et al., 2022) and Object Motion Guided Human Motion Synthesis (Li et al., 2023) capture human body motion with dynamic objects (e.g.,, suitcases) without hand grasping. Hand-only datasets (Chao et al., 2021; Zhang et al., 2021) focus on finger grasping sequences without body motions. GRAB (Taheri et al., 2020) and ARCTIC (Fan et al., 2023) capture whole-body manipulation with hand articulation but limited action diversity. ParaHome (Kim et al., 2025) captures indoor sequences with both locomotion and finger-level manipulations. For kinematic synthesis, IMoS (Ghosh et al., 2023) generates intent-driven full-body motions for object interactions, and FORCE (Zhang et al., 2024) introduces physics-aware synthesis with force modeling.

2.3 Physics-Based Character Control

DeepMimic (Peng et al., 2018) demonstrated that deep RL can successfully track reference motion in physics simulation. Adversarial reward settings (Peng et al., 2021; Xu & Karamouzas, 2021; Zhang et al., 2025) introduced stylized control without explicit reference tracking, while latent-based methods (Peng et al., 2022; Dou et al., 2023; Tessler et al., 2023) learn reusable skill-embeddings for compositional task execution. Recent work scaled motion tracking with large-scale datasets (Mahmood et al., 2019) to learn single policies for real-time tracking (Luo et al., 2023), with adaptations for language control (Juravsky et al., 2022; 2024) and multi-task composition via masked observations (Tessler et al., 2024). For physics-based manipulation, D-Grasp (Christen et al., 2022) synthesizes dynamic grasps for hand-object interactions, and PhysHOI (Wang et al., 2023) tackles physics-based imitation of dynamic human-object interaction. For scene-level interactions, InterPhys (Hassan et al., 2023) synthesizes physical character-scene interactions with contact handling, UniHSI (Xiao et al., 2024) performs HSI using chain-of-contacts with LLM planning, and TokenHSI (Pan et al., 2025) demonstrates multi-task composition with task tokenization.

2.4 Behavior Foundation Models

A behavior foundation model (Tirinzoni et al., 2025) refers to a universal policy that generalizes across diverse motion styles trained with large-scale datasets. MoConVQ (Yao et al., 2023) uses discrete representations as compact motion priors that interface with language models. PULSE (Luo et al., 2024) trains a universal policy with CVAE aimed for adaptation to diverse downstream tasks. More recent work utilizes task-specific tokens as flexible representations to adapt general tracking policies (Tessler et al., 2024; Vainshtein et al., 2025). Composite motion learning (Xu et al., 2023) enables task-aware skill composition for complex behaviors. Based on these advances, we aim to integrate locomotion datasets (Mahmood et al., 2019) and whole-body manipulation datasets (Taheri et al., 2020) through a staged training strategy that learns a general motion prior and then steer the trained model to perform various tasks.

3 PRELIMINARIES

Given a 3D scene with static structures such as furniture and movable objects, our goal is to output a joint-level action target that interacts with the scene and manipulates objects within it. The resulting motion coordinates locomotion, scene contact, and hand-object interaction to reach task goals and to move between selected skills, without a large-scale reference motion dataset that covers the entire demonstration sequence.

3.1 FORMULATION

We formulate character control as a goal-conditioned reinforcement learning problem. Let $s_t \in \mathcal{S}$ denote the simulation state, including the character and its environment, $g_t \in \mathcal{G}$ the task condition, and $a_t \in \mathcal{A}$ the joint-level control action. The task condition specifies the active interaction and its targets, and may change as the state progresses. Physical transitions follow $P(s_{t+1} \mid s_t, a_t)$, while task conditions are updated according to the each task's progression rules.

At each control step, the policy selects an action $a_t \sim \pi(\cdot \mid s_t, g_t)$ and receives a task-dependent reward $r_t = R(s_t, a_t, g_t)$. Our objective is to learn a policy that maximizes the expected discounted return over an episode of T control steps:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi} \left[\sum_{t=0}^{T-1} \gamma^t R(s_t, a_t, g_t) \right], \tag{1}$$

where $\gamma \in [0, 1)$ is the discount factor and the expectation is over task and initial-state sampling and the resulting trajectories.

We use an SMPL-X-based humanoid with 52 joints (Pavlakos et al., 2019), including a floating root, following (Luo et al., 2023). Each of the 51 non-root joints has three actuated degrees of freedom, giving a joint-level action space $\mathcal{A} \subset \mathbb{R}^{153}$. The action a_t specifies joint position targets tracked by proportional–derivative (PD) controllers; the root's global translation and rotation evolve through the simulated dynamics. Here, π denotes the complete controller producing these joint-level actions; its motion-prior-based parameterization is given in Eq. 2.

4 METHOD

Our framework (Fig. 2) brings scene interaction and dexterous manipulation into a shared control formulation, with task conditions specifying spatial goals and phase-dependent interaction requirements. For scene interaction, these conditions specify movement or contact goals, such as a pelvis target on a chair for sitting. For manipulation, they specify the desired object pose, active hands, object-relative grasp targets, and whether each hand should maintain or release its grasp. To leverage a learned whole-body motion prior for scene interaction and manipulation, we build on the pretrained prior and decoder of PULSE (Luo et al., 2024). Given proprioceptive observations and the previous action, denoted together by x_t, the prior predicts a distribution over latent codes that the decoder maps to joint-level action targets. Its mean, $\mu_\theta(x_t)$, provides the state-conditioned base code. A transformer uses character, environment, and task observations to predict a Gaussian latent residual Δz_t, which adjusts the prior mean before decoding:

$$z_t = \mu_\theta(x_t) + \Delta z_t, \qquad a_t = D_\theta(x_t, z_t). \tag{2}$$

The prior and decoder remain fixed while the task-conditioned policy is trained.

4.1 UNIFIED INTERACTION TASK REPRESENTATION

Task tokens. A shared policy must accommodate different interaction conditions and combine them when scene interaction and manipulation are required simultaneously. We therefore organize observations and task conditions into groups according to their roles and map each group o_i to a fixed-dimensional token $h_i = T_i(o_i)$ using its corresponding tokenizer T_i, following TokenHSI (Pan et al., 2025). Shared tokens encode the character's proprioceptive state, previous action, motion-prior mean, and local scene heightmap. Scene-interaction tokens encode spatial goals and the poses and bounding boxes of interacting objects. Manipulation tokens encode the manipulated object's

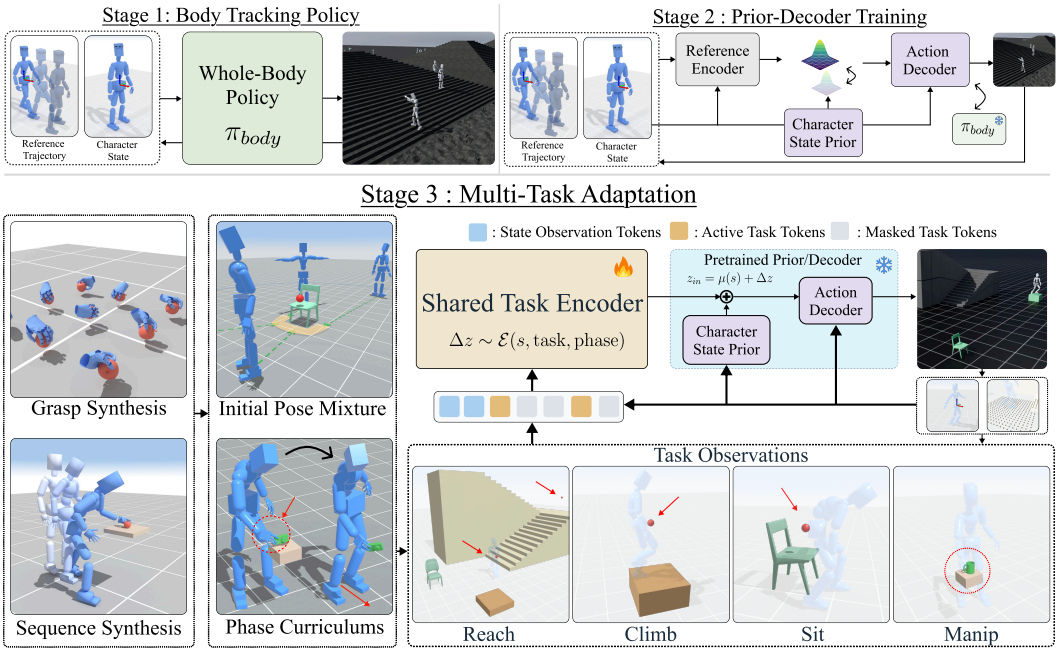

Figure 2: **Overview.** We first learn a motion prior and decoder from a whole-body tracking policy, then adapt a shared task encoder for scene interaction and manipulation tasks.

state, geometry, and target pose. Per-hand tokens additionally encode the desired grasp as an object-relative wrist pose and the corresponding wrist target expressed relative to the character.

Masked task conditioning. Each token h_i is associated with a binary mask m_i, where $m_i = 0$ excludes the token from attention. Shared tokens remain active, while task tokens are enabled according to the current task. For example, sitting while holding an object activates scene-interaction tokens together with object and grasp tokens. The transformer defines a distribution over latent residuals conditioned on the token sequence and its corresponding masks:

$$\Delta z_t \sim \mathcal{E}([h_1, \ldots, h_N], [m_1, \ldots, m_N]) . \tag{3}$$

4.2 Goal-Conditioned Loco-Manipulation Curriculum

The task representation specifies the desired interaction, but learning to map it to the actual action requires known coordinated body motion and precise contact formation. Reference motions provide a dense guidance for these behaviors, but do not specify how to achieve the same goal from different initial states or object placements. We therefore combined reference-guided and goal-driven learning, varying the scope of reference guidance according to the task.

Goal-conditioned scene interaction. We formulate standing up from a chair, descending from a box, and traversing stairs as Reach tasks, using intermediate waypoints as reward where needed. For Sit and Climb-up, we jointly train the same policy in reference and exploration environments. Reference environments are initialized relative to the demonstration and provide tracking supervision for precise interaction. Exploration environments instead sample starting positions at varying distances and directions around the furniture, requiring the character to approach and complete the same interaction from states beyond the demonstrated approach.

Pick-move-place manipulation. We train pick-move-place using an end-target object pose $G^*_{o,t}$ and grasp/release conditions, rather than the object trajectories used by OmniGrasp (Luo et al., 2025). The object goal alone does not determine how the hands should grasp the object or when they should disengage. We therefore provide an object-relative grasp wrist pose $G^{h,*}_{w|o}$ for each

active hand $h \in \{L, R\}$. Given the current object pose $G_{o,t}$, the corresponding world-space wrist target is

$$G^{h,*}_{w,t} = G_{o,t} G^{h,*}_{w|o}. \tag{4}$$

This object-relative specification preserves the desired hand and object alignment across object placements, and the resulting wrist targets are expressed relative to the character for policy input. Per-hand binary conditions $b^h_t \in \{0, 1\}$ indicate whether each active hand should retain or release its grasp. These targets and per-hand conditions form the manipulation component of g_t and are updated as execution progresses through grasp formation, lifting, transport, placement, and release. During placement, we sample a dwell interval in which the hands retain their grasp before the release condition is activated, distinguishing arrival at the target pose from completion of placement and release. Reference motion guides initial body coordination and wrist to finger alignment during the close range grasp formation. At the reference lift time, we remove full-body tracking and continue with task-specific goals, while actual lift completion determines when transport begins.

4.3 Cross-Task Curriculum

Joint training of individual and sequential tasks. We jointly train individual scene and manipulation tasks with selected manipulation-to-scene sequences, in which grasp-and-lift is followed by sitting, climbing, or reaching. Once lifting is complete, we activate the destination scene goal and its task tokens without resetting the character or object state. The subsequent task therefore begins from the state produced by lifting, rather than an independently sampled initialization. Without a reference motion spanning the sequence, the policy explores how to continue into the scene task under task rewards.

Maintaining an interaction across tasks. The scene task must be performed while preserving the grasp established during lifting. We therefore keep object and grasp tokens enabled alongside the scene-task tokens, with grasp-retention conditions $b^h_t = 1$ for the active hands. We define the carried object's target $G^*_{o,t}$ from its pose relative to the character's heading frame at handover and update it with the character's motion. Placement and release remain inactive during this segment. These conditions jointly guide scene interaction and grasp retention.

5 Experiments

We evaluate individual scene-interaction and manipulation capabilities, followed by their sequential and simultaneous execution. For scene interaction, we compare our shared policy with per-task specialists and an adapted MaskedMimic baseline. We also assess the effect of synthetic data augmentation through reference-conditioned lifting with MimicManipulator, and evaluate our policy on goal-conditioned pick-move-place tasks without prescribed object trajectories. Finally, we qualitatively demonstrate manipulation-to-scene transitions and scene interaction while maintaining an object grasp.

5.1 Setups

Datasets. We use AMASS (Mahmood et al., 2019) and GRAB (Taheri et al., 2020) dataset throughout the tasks. To enrich finger articulation, we augment AMASS motions with grasping segments sampled from GRAB. For HOI training, we combine mocap references, including their object poses, with synthetic sequences generated using Dexonomy (Chen et al., 2025) and Kimodo (Rempe et al., 2026). We filter the synthetic sequences to remove unnatural whole-body motions during object manipulation. For reference-conditioned lifting and holding, we use grasp-and-lift segments from GRAB, with data preparation and splits detailed in Appendix A. For scene interactions, we use motion data, human-object interaction intervals, and object geometry provided by TokenHSI (Pan et al., 2025).

Simulation setups. We use the ProtoMotions (Tessler et al., 2024) framework with IsaacLab (Mittal et al., 2025). Policies operate at 30 Hz, with physics frequency at 120 Hz. During tracker training and prior distillation, we apply random pushes with randomized wrist and finger masses to improve

robustness for downstream object manipulations as well as complex terrains with pyramidal-stairs and hills for the ease of the scene interaction. We include further details for training in appendix.

Policy architecture. Following PULSE (Luo et al., 2024), we use MLPs for the motion prior and decoder. Both remain frozen during task-policy training. The task encoder is a four-layer transformer with four attention heads and 512-dimensional tokens.

Baselines. For scene interaction, we adapt MaskedMimic (Tessler et al., 2024) to our task environments by adding observations of surrounding objects and distilling a tracking teacher policy. We also train a separate instance of our task policy for each scene task, denoted as per-task specialists to assess the quality of sharing a policy across tasks. For manipulation, we implement MimicManipulator (Tessler et al., 2025) which is a reference trajectory conditioned tracker to evaluate synthetic data augmentation comparing training on mocap references along with training on mocap and synthetic references. We additionally use its reference-conditioned lifting performance to gauge the picking performance of our goal-conditioned pick-move-place policy.

5.2 Metrics

For scene-interaction tasks, we report task success rate (SR, %) and target position error (cm). All distances below are measured between the interacting targets and the goal position.

Success rate. *Sit.* Each chair has a predefined sitting target. Success requires the pelvis to remain within 10 cm of this target for 60 consecutive control steps. *Climb-up.* Success is achieved when the pelvis is within 30 cm of the target above the box center. *Stand-up.* The character starts seated on a chair. Success requires the pelvis to be at least 85 cm above the ground and simultaneously within 30 cm of the goal before timeout. *Climb-down.* The character starts on top of a box. Success requires the pelvis to reach within 30 cm of the ground-level goal without falling. Neither Stand-up nor Climb-down requires a holding interval at the end. *Stair.* The character starts on the ground near the staircase. Success requires first reaching within 30 cm of the middle waypoint, then remaining within 30 cm of the final waypoint. The reported Stair SR measures completion of this full sequence.

Target-position error. For Sit and Stair tasks, we average the error over the first successful holding interval, using the seat target and the final goal point. For Climb-up, Stand-up, and Climb-down tasks, we measure the error at the first successful control step. We then average these per-episode errors over successful episodes with equal weights.

HOI reference lifting. We evaluate reference-conditioned lifting with success rate, mean per-joint position error (MPJPE), and tracking length. Each reference is truncated at the first frame where the object rises at least 10 cm above its initial height. A trial succeeds if tracking reaches this endpoint without any humanoid rigid-body position error exceeding 0.5 m or object position error exceeding 0.25 m, and without numerical failure. MPJPE measures world-space position error over all humanoid rigid bodies, averaged over valid tracking frames within each trial and then across trials; terminal failure frames are excluded Tracking length measures elapsed control time until the first tracking failure or reference completion, including the terminal detection step, averaged across trials.

Pick-move-place. We report pick, move, and release success rates at each phase. Pick and move require hand object contact and object position errors below 3 cm and 5 cm. Release is commanded after a 2 s dwell in the placement target region and succeeds after 15 contact-free steps within 5 cm of the target, using a 1 N contact threshold. Contact loss for 20 consecutive steps before the command or continued contact for 15 steps afterward causes failure. Release error is averaged over the successful 15-step window and across successful episodes.

5.3 Quantitative Results

Scene-Interaction Task Comparison. Table 1 compares our shared policy and per-task specialists with MaskedMimic. Both variants achieve higher success rates and lower target-position errors than

Method	Sit		Climb-up		Reach(Stand-up)		Reach(Climb-down)		Reach(Stair)	
	SR (%)	Error [cm]	SR (%)	Error [cm]	SR (%)	Error [cm]	SR (%)	Error [cm]	SR (%)	Error [cm]
MaskedMimic	44.27	8.36	93.23	28.72	72.40	29.55	83.33	29.50	7.81	29.07
Ours (Per-Task Specialists)	**99.48**	**5.82**	**100.00**	27.74	98.31	**23.93**	**100.00**	**26.79**	54.41	17.59
Ours	93.22	6.68	**100.00**	**27.73**	**100.00**	27.62	**100.00**	27.38	**69.23**	**11.58**

Table 1: **Scene interaction tasks.** For each task, our per-task policy outperforms the baseline while our task-integrated policy works on par with each expert policy at each task.

(a) HOI Lifting Task

Split	Training data	SR (%) ↑	MPJPE [cm] ↓	Length [s] ↑
Train	GRAB Only	**93.66**	**2.42**	**2.26**
	GRAB + Synthetic	83.41	3.04	2.25
Test	GRAB Only	56.67	3.29	1.78
	GRAB + Synthetic	**90.00**	**3.11**	**1.99**

(b) Pick-Move-Place Task

Split	Pick SR (%) ↑	Move SR (%) ↑	Release SR (%) ↑	Release Error [cm] ↓
Train	81.02	77.29	63.39	2.34
Test	84.80	75.20	64.80	2.49
Unseen	47.90	40.25	25.19	2.87

Table 2: **Reference tracking and pick-move-place evaluation.** (a) Effect of synthetic training data augmented to mocap data and trained on the tracking policy. (b) Picking, moving, and releasing the given object, we measure success rates at each conditioned interact phase and measure release error of the object at the goal pose.

MaskedMimic across all five tasks. The shared policy achieves 100.00% SR on Climb-up, Stand-up, and Climb-down, and 93.22% on Sit. Stair remains the most challenging task, with an SR of 69.23%, compared with 7.81% for MaskedMimic. Compared with per-task specialists, the shared policy improves Stair SR from 54.41% to 69.23% and reduces error among successful episodes from 17.59 to 11.58 cm. It also increases Stand-up SR from 98.31% to 100.00% and matches specialist SRs on Climb-up and Climb-down, while Sit SR decreases from 99.48% to 93.22%. These results show that a shared policy supports multiple scene interactions with task-dependent trade-offs, rather than a loss in performance relative to each specialized policies.

Data Augmentation Justification. In Table 2(a), we train MimicManipulator on the GRAB dataset and its augmented version. On train dataset, the model trained with only GRAB dataset show higher SR with lower MPJPE compared to the reference used. However on test set, the result differs by large margin compared to train set result, substantially improving the SR from 56.67% to 90.00%. MPJPE on test set also reduces from 3.29 to 3.11 cm and the average tracking length increases from 1.78 to 1.99 s. Such train-test gap was also observed in MimicManipulator (Tessler et al., 2025), and we believe the data augmentation can ease such discrepancy. Overall, the result suggests that the additional diversity introduced by the synthetic sequences reduces overfitting to the training references and improves generalization to unseen motions with known object identities.

Pick-Move-Place Evaluation. Table 2(b) reports cumulative pick, move, and place success rates, each normalized by the total number of evaluation episodes. Our policy achieves a training-set Pick SR of 81.02%, compared with 83.41% SR for the reference-conditioned tracker trained with augmented data. This comparison provides context rather than a direct comparison, since the tracker measures reference completion whereas pick requires maintained contact and proximity to the lift target, as well as difference in conditioning of trajectory.

The policy completes the full pick-move-place sequence in 63.39% and 64.80% of train and test episodes. On unseen objects, pick, move, and release success rates decrease to 47.90%, 40.25%, and 25.19%, showing partial transfer with a substantial drop in overall completion. Among successful release episodes, mean placement errors remain between 2.34 and 2.87 cm across splits, indicating similar placement precision conditional on successful completion.

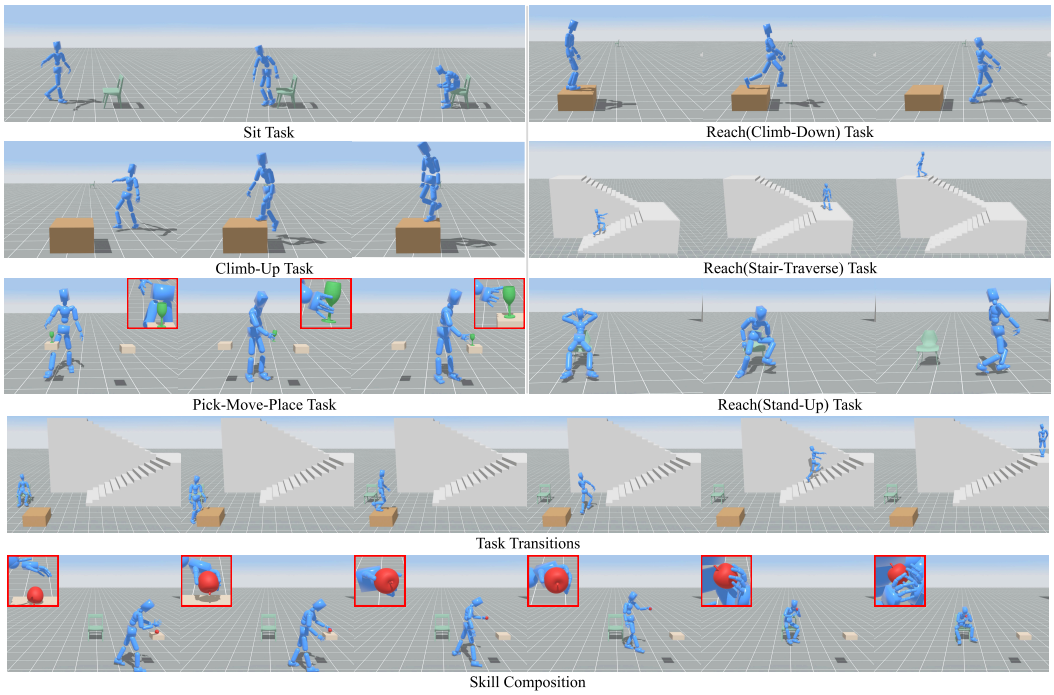

Figure 3: **Deployment of each task.** Our integrated policy demonstrates multiple tasks, encompassing scene-interaction to object manipulation. The model also covers transitioning between the tasks with compositing manipulation and scene-tasks.

5.4 Qualitative Results

Fig. 3 qualitatively illustrates our policy performing multiple tasks in the environment containing both static structures and manipulable objects. The character navigates toward designated goals, grasps and transports objects to target poses, and positions its body for interactions with the scene. By specifying task goals and masks, we compose tasks learned during training into sequences of navigation, object manipulation, and body-scene interaction. The illustrated rollouts exhibit smooth transitions between tasks while maintaining balance and coordinated full-body motion, consistent with the natural motion encouraged by the motion prior.

6 Conclusion

In this work, we present a framework that unifies scene interaction and dexterous manipulation through a shared policy formulation. A task-conditioned encoder predicts residual actions in a latent space defined by a frozen motion prior and decoder, while task and phase masks accommodate interaction-specific goals and observations switched at each defined task-phase. Partial reference guidance and task-driven exploration support individual interactions, while cross-task training enables manipulation-to-scene transitions with continued grasp pose.

Our experiments demonstrate scene interaction and object manipulation under varied initial conditions, and qualitative results illustrate coordinated full-body motion across task transitions. Among successful trials, placement precision remains comparable across evaluation splits. These results show generalization to unseen object instances, while highlighting reliable interaction completion as a remaining challenge.

Future Work. Future work could incorporate visual and language-based task conditioning to support higher-level interaction specifications. Extending the range of trained transitions and improving robustness to unseen objects could further support longer and more diverse task sequences. Another interesting direction is to investigate how the learned motion prior and interaction policies can be adapted to real-world humanoids for the actual deployment.

AI USE STATEMENT

In this work, we used generative AI tools for generate synthetic data sets, assist with translation, design or provide feedback on research methodology or experiments, and implement methods. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, propose or refine hypotheses, clean and reformat dataset, support qualitative and thematic data analysis, interpret results and formulate mathematical claims, provide critical ingredients for proving mathematical claims are not applicable to this work. Additionally, we used generative AI tools for suggest experimental parameters. We have reviewed all AI-assisted work. We checked LLM-generated research ideas for potential plagiarism through a manual literature survey. LLM-generated code was verified and tested for correctness by the author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 describes the policy formulation and training curricula. Section 5 specifies the data sources, simulation setup, baseline implementations, and evaluation metrics. Appendix A provides further details on reference preparation and synthetic data generation, task rewards and termination conditions, state handover between tasks, network architecture, and optimization settings.

REFERENCES

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Behave: Dataset and method for tracking human object interactions, 2022. URL https://arxiv.org/abs/2204.06950.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A Benchmark for Capturing Hand Grasping of Objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9040–9049, Nashville, TN, USA, June 2021. IEEE. ISBN 978-1-66544-509-2. doi: 10.1109/CVPR46437.2021.00893.

Jiayi Chen, Yubin Ke, Lin Peng, and He Wang. Dexonomy: Synthesizing all dexterous grasp types in a grasp taxonomy, 2025. URL https://arxiv.org/abs/2504.18829.

Sammy Christen, Muhammed Kocabas, Emre Aksan, Jemin Hwangbo, Jie Song, and Otmar Hilliges. D-Grasp: Physically Plausible Dynamic Grasp Synthesis for Hand-Object Interactions, March 2022.

Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C$\cdot$ASE: Learning Conditional Adversarial Skill Embeddings for Physics-based Characters. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, December 2023. doi: 10.1145/3610548.3618205.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation, April 2023.

Anindita Ghosh, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Philipp Slusallek. IMos: Intent-Driven Full-Body Motion Synthesis for Human-Object Interactions, February 2023.

Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Yunus Saracoglu, Torsten Sattler, and Gerard Pons-Moll. Interaction Replica: Tracking Human-Object Interaction and Scene Changes From Human Motion. https://arxiv.org/abs/2205.02830v4, May 2022.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic Scene-Aware Motion Prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11354–11364, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.01118.

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. Synthesizing Physical Character-Scene Interactions, February 2023.

Nan Jiang, Zimo He, Zi Wang, Hongjie Li, Yixin Chen, Siyuan Huang, and Yixin Zhu. Autonomous Character-Scene Interaction Synthesis from Text Instruction, October 2024a.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. Scaling Up Dynamic Human-Scene Interaction Modeling, May 2024b.

Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. PADL: Language-Directed Physics-Based Character Control. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–9, November 2022. doi: 10.1145/3550469.3555391.

Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. SuperPADL: Scaling Language-Directed Physics-Based Control with Progressive Supervised Distillation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24*, pp. 1–11, July 2024. doi: 10.1145/3641519.3657492.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling of Human-Object Interactions, January 2025.

Jiye Lee and Hanbyul Joo. Locomotion-Action-Manipulation: Synthesizing Human-Scene Interactions in Complex 3D Environments, September 2023.

Jiaman Li, Jiajun Wu, and C. Karen Liu. Object motion guided human motion synthesis, 2023. URL https://arxiv.org/abs/2309.16237.

Zhengyi Luo, Jinkun Cao, Alexander Winkler, Kris Kitani, and Weipeng Xu. Perpetual Humanoid Control for Real-time Simulated Avatars, September 2023.

Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal Humanoid Motion Representations for Physics-Based Control, April 2024.

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. Omnigrasp: Grasping Diverse Objects with Simulated Humanoids, May 2025.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pp. 5442–5451, October 2019.

Mayank Mittal, Pascal Roth, James Tigue, Antoine Richard, Octi Zhang, Peter Du, Antonio Serrano-Muñoz, Xinjie Yao, René Zurbrügg, Nikita Rudin, Lukasz Wawrzyniak, Milad Rakhsha, Alain Denzler, Eric Heiden, Ales Borovicka, Ossama Ahmed, Iretiayo Akinola, Abrar Anwar, Mark T. Carlson, Ji Yuan Feng, Animesh Garg, Renato Gasoto, Lionel Gulich, Yijie Guo, M. Gussert, Alex Hansen, Mihir Kulkarni, Chenran Li, Wei Liu, Viktor Makoviychuk, Grzegorz Malczyk, Hammad Mazhar, Masoud Moghani, Adithyavairavan Murali, Michael Noseworthy, Alexander Poddubny, Nathan Ratliff, Welf Rehberg, Clemens Schwarke, Ritvik Singh, James Latham Smith, Bingjie Tang, Ruchik Thaker, Matthew Trepte, Karl Van Wyk, Fangzhou Yu, Alex Millane, Vikram Ramasamy, Remo Steiner, Sangeeta Subramanian, Clemens Volk, CY Chen, Neel Jawale, Ashwin Varghese Kuruttukulam, Michael A. Lin, Ajay Mandlekar, Karsten Patzwaldt, John Welsh, Huihua Zhao, Fatima Anes, Jean-Francois Lafleche, Nicolas Moënne-Loccoz, Soowan Park, Rob Stepinski, Dirk Van Gelder, Chris Amevor, Jan Carius, Jumyung Chang, Anka He Chen, Pablo de Heras Ciechomski, Gilles Daviet, Mohammad Mohajerani, Julia von Muralt, Viktor Reutskyy, Michael Sauter, Simon Schirm, Eric L. Shi, Pierre Terdiman, Kenny Vilella, Tobias Widmer, Gordon Yeoman, Tiffany Chen, Sergey Grizan, Cathy Li, Lotus Li, Connor Smith, Rafael Wiltz, Kostas Alexis, Yan Chang, David Chu, Linxi "Jim" Fan, Farbod Farshidian, Ankur Handa, Spencer Huang, Marco Hutter, Yashraj Narang, Soha Pouya, Shiwei Sheng, Yuke Zhu, Miles Macklin, Adam Moravanszky, Philipp Reist, Yunrong Guo, David Hoeller, and Gavriel State. Isaac lab: A gpu-accelerated simulation framework for multi-modal robot learning. *arXiv preprint arXiv:2511.04831*, 2025. URL https://arxiv.org/abs/2511.04831.

Liang Pan, Zeshi Yang, Zhiyang Dou, Wenjia Wang, Buzhen Huang, Bo Dai, Taku Komura, and Jingbo Wang. TokenHSI: Unified Synthesis of Physical Human-Scene Interactions through Task Tokenization, April 2025.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. 2019.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. DeepMimic: Example-Guided Deep Reinforcement Learning of Physics-Based Character Skills. *ACM Transactions on Graphics*, 37(4):1–14, August 2018. ISSN 0730-0301, 1557-7368. doi: 10.1145/3197517.3201311.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Transactions on Graphics*, 40 (4):1–20, August 2021. ISSN 0730-0301, 1557-7368. doi: 10.1145/3450626.3459670.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. ASE: Large-Scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters. *ACM Transactions on Graphics*, 41(4):1–17, July 2022. ISSN 0730-0301, 1557-7368. doi: 10.1145/3528223.3530110.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation, 2026. URL https://arxiv.org/abs/2603.15546.

Sebastian Starke, He Zhang, Taku Komura, and Jun Saito. Neural state machine for character-scene interactions. *ACM Trans. Graph.*, 38(6):209:1–209:14, 2019. ISSN 0730-0301. doi: 10.1145/3355089.3356505.

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV*, pp. 581–600, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58547-1. doi: 10.1007/978-3-030-58548-8_34.

Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. CALM: Conditional Adversarial Latent Models for Directable Virtual Characters. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*, pp. 1–9, July 2023. doi: 10.1145/3588432.3591541.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. MaskedMimic: Unified Physics-Based Character Control Through Masked Motion Inpainting, September 2024.

Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. MaskedManipulator: Versatile Whole-Body Manipulation, September 2025.

Andrea Tirinzoni, Ahmed Touati, Jesse Farebrother, Mateusz Guzek, Anssi Kanervisto, Yingchen Xu, Alessandro Lazaric, and Matteo Pirotta. Zero-Shot Whole-Body Humanoid Control via Behavioral Foundation Models, April 2025.

Ron Vainshtein, Zohar Rimon, Shie Mannor, and Chen Tessler. Task Tokens: A Flexible Approach to Adapting Behavior Foundation Models, March 2025.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. PhysHOI: Physics-Based Imitation of Dynamic Human-Object Interaction, December 2023.

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. Unified Human-Scene Interaction via Prompted Chain-of-Contacts, November 2024.

Pei Xu and Ioannis Karamouzas. A GAN-Like Approach for Physics-Based Imitation Learning and Interactive Character Control. *Proceedings of the ACM on Computer Graphics and Interactive Techniques*, 4(3):1–22, September 2021. ISSN 2577-6193. doi: 10.1145/3480148.

Pei Xu, Xiumin Shang, Victor Zordan, and Ioannis Karamouzas. Composite Motion Learning with Task Control. *ACM Transactions on Graphics*, 42(4):1–16, August 2023. ISSN 0730-0301, 1557-7368. doi: 10.1145/3592447.

Heyuan Yao, Zhenhua Song, Yuyang Zhou, Tenglong Ao, Baoquan Chen, and Libin Liu. MoConVQ: Unified Physics-Based Motion Control via Scalable Discrete Representations, December 2023.

He Zhang, Yuting Ye, Takaaki Shiratori, and Taku Komura. ManipNet: Neural manipulation synthesis with a hand-object spatial representation. *ACM Transactions on Graphics*, 40(4):1–14, August 2021. ISSN 0730-0301, 1557-7368. doi: 10.1145/3450626.3459830.

Xiaohan Zhang, Bharat Lal Bhatnagar, Sebastian Starke, Ilya Petrov, Vladimir Guzov, Helisa Dhamo, Eduardo Pérez-Pellitero, and Gerard Pons-Moll. FORCE: Physics-aware Human-object Interaction, December 2024.

Ziyu Zhang, Sergey Bashkirov, Dun Yang, Yi Shi, Michael Taylor, and Xue Bin Peng. Physics-Based Motion Imitation with Adversarial Differential Discriminators. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, December 2025. doi: 10.1145/3757377.3763819.

A Training Details

A.1 Data Preparation

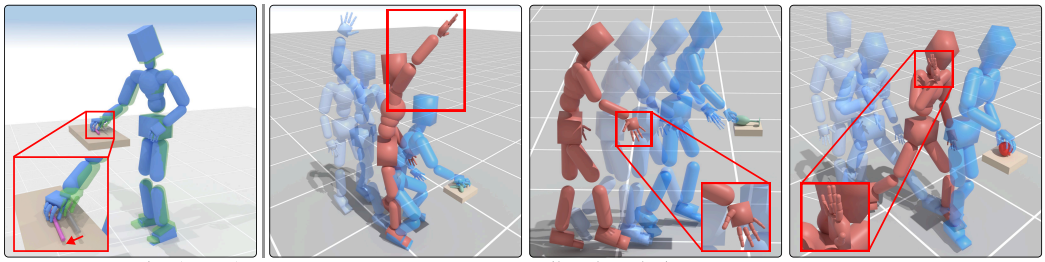

Figure 4: **Reference preparation.** Left: contact-based object retargeting for mocap references. Right: rejected synthetic sequences with unnatural arm motion or invalid joint configurations, shown in red.

Data composition and splits. Manipulation references comprise mocap sequences from GRAB (Taheri et al., 2020) and synthetic sequences generated using Dexonomy (Chen et al., 2025) and Kimodo (Rempe et al., 2026). The mocap subset contains six objects: apple, camera, cup, wine-glass, water bottle, and coffeemug. Subjects from 1 to 9 are used for training and subject 10 for testing. The combined dataset contains 2,618 training sequences (110 mocap and 2,508 synthetic) and 303 test sequences (18 mocap and 285 synthetic).

Contact-based retargeting. We retarget mocap objects to the neutral humanoid using contact correspondences, following a procedure similar to MaskedManipulator (Tessler et al., 2025). For each frame, we record the contacting body parts and corresponding object vertices. The wrist displacement between the original and neutral characters provides an initial object-position correction; object placement is then refined by minimizing distances between corresponding contact points. Contacts are identified through distance queries between the object mesh and surface points sampled from the character's transformed geometry primitives. Figure 4 illustrates the resulting alignment.

Synthetic reference generation. We generate left- and right-hand MANO grasps from 14 predefined grasp types. Surface-based hand initializations are refined jointly in wrist pose and finger articulation under contact-preservation constraints. Grasps that penetrate the object or ground are rejected, and the remaining grasps are validated by an object-lifting test in MuJoCo.

Each accepted grasp specifies an object-relative wrist pose. For a sampled object placement, the world-space wrist target follows Eq. 4. After aligning the MANO and SMPL-X hand frames, upper-body inverse kinematics provides reachable end-effector constraints. Kimodo generates from 1.5 to 2.0 s approach-and-grasp sequences at 30 Hz, conditioned on these constraints and an initial pose sampled from the mocap data. Finger articulation is transferred from the grasp, with closure over the final 10 frame intervals and a subsequent 5 s terminal hold.

Sequence filtering. Retained sequences have wrist-target error below 1 cm and downward wrist displacement near grasp acquisition below 5 cm. We additionally reject sequences with invalid terminal contacts, initial body object interferences, inverted hand orientations, or pre-contact wrist motion above the shoulder. Examples of rejected sequences appear in Fig. 4.

A.2 Task Objectives and Initialization

The observation and action representations follow Sec. 4.1 and Eq. 2. Below we specify training rewards, initialization, and termination conditions; evaluation criteria are defined separately in Sec. 5.2.

Sit and Climb-up. Both tasks combine reference-state initialization with goal-directed exploration. Reference-initialized environments use body-tracking rewards without providing an explicit reference trajectory as policy input. Exploration environments use approach and target-proximity rewards, with an additional foot-height term for climbing.

For Sit, the pelvis target is extracted from the terminal reference segment in the chair frame. For polar starts around the chair, the interaction reward is

$$r_{\text{sit}} = \exp[-1.5(1 - \cos \Delta\psi)] \exp(-10d^2),$$

where d denotes pelvis-to-target distance and $\Delta\psi$ the horizontal heading error. This reward is multiplied by the progress factor $(0.5 + p_t)$ defined in Sec. A.3. Completion requires $d < 0.1$ m and heading error below 20° for 60 consecutive control steps. Climb-up uses a target above the box surface, with a training completion threshold of 0.15 m for 60 consecutive steps. Reference-aligned episodes terminate when tracking error exceeds 0.5 m; default-pose starts receive a 60-step grace period.

Reach. Reach variants comprise standing from a chair, descending from a box, free walking, and stair traversal, with corresponding initial poses and support geometry. Stand-up and Climb-down goals are sampled from 1 to 5 m and from 1 to 3 m from the initial position, respectively. Rewards combine goal-directed velocity, facing, and proximity, supplemented by height recovery for standing and height-progress and power terms for stair traversal. Stair starts and goals are randomized. Intermediate waypoints advance at horizontal and vertical distances below 0.3 m and 0.4 m, respectively; final-waypoint completion requires horizontal distance below 0.15 m for 60 consecutive steps.

Reference-conditioned lifting. The HOI tracker observes body and object targets 1 and 30 frames ahead. Mocap references end when the object rises 10 cm above its initial height, whereas synthetic references retain the generated grasp-and-hold segment. The tracking reward is the product of body-tracking, local hand-tracking, object-interaction, and power-regularization terms. Training terminates when body or object tracking error exceeds 0.25 m or when reference-dependent contact conditions are violated.

Pick-move-place. Episodes begin before the reference's first contact, with perturbed root position and heading. Let r_{track}, r_w, r_f, r_n, r_c, r_h, and r_g denote body tracking, wrist alignment, finger articulation, grasp-normal alignment, surface-contact proximity, body height regularization, and object-goal tracking, respectively. The grasp and lift rewards are

$$\begin{aligned} r_{\text{grasp}} &= r_{\text{track}}(0.5 r_w r_f r_n + 0.5 r_c), \\ r_{\text{lift}} &= 0.25 r_f r_w r_h + 0.5 r_c + 0.25 r_g. \end{aligned}$$

The switch to r_{lift} removes full-body tracking at the reference lift timestamp, independently of physical lift completion. Transport begins only when the object reaches within 3 cm of the lift goal. Its reward is $0.5 r_{\text{hand}} r_h + 0.5 r_g$, where r_{hand} combines finger, wrist, and surface-contact terms. Transport goals are sampled from 0.75 to 1.2 m from the transport starting point, with a placement-time signal sampled from 0.5 to 3 s.

Release and continued-holding episodes are sampled with equal probability. Release requires 15 consecutive steps without contact at a 1 N threshold; continued holding requires 90 steps at the object goal. Episodes terminate on missed lift deadlines, contact failures, or excessive wrist deviation during transport.

A.3 Cross-Task Training

Task sampling and state handover. We allocate 30% of environments to scene tasks, 35% to manipulation, and 35% to lifting followed by scene interaction. Within the latter group, Sit and Climb-up are each sampled with probability $1/3$, and free walking, Climb-down, and Stair with probability $1/9$ each. Upon lift completion, the destination task begins without resetting body pose, velocities, or object state. Goal-progress measurement restarts from the handover state, and the object's pose relative to the character's heading frame defines the moving carry target. Object and grasp tokens remain active; placement and release are disabled.

Joint objective and termination. Scene execution and object retention are trained jointly using

$$r_{\text{transition}} = (0.5 + p_t) r_{\text{scene}} + \lambda r_{\text{carry}},$$
$$r_{\text{carry}} = 0.5 r_c r_f + 0.5 r_g,$$

where r_{scene} is the destination-task reward and $\lambda = 0.1$ for Climb-up and 0.5 otherwise. For eligible steps, $p_t = \min(30\Delta d_t, 1)$ measures improvement $\Delta d_t \geq 0$ in the minimum distance reached during the current stage; $p_t = 1$ in the final goal region and zero on ineligible steps. During carrying, eligibility requires maintained hand contact and no scene-task failure. Each active hand must register contact above 1 N on at least one required contact body. Episodes terminate on contact loss, scene-task failure, or timeout. Completion requires destination-task success under these constraints, without the additional completion bonus used for standalone manipulation.

A.4 Architecture and Optimization

Table 3 summarizes the task-policy architecture and PPO settings. The motion prior and decoder remain frozen, and joint task training uses neither teacher-encoder supervision nor its KL loss. Observations are normalized and clipped to ± 5. We use moving-average advantage normalization and clipped value loss; actor updates are skipped when the minibatch clipping fraction exceeds 0.85.

Table 3: **Task-policy architecture and optimization settings.**

Setting	Value
Tokenizer MLP	2 hidden layers, 256 units, SiLU
Token dimension	512
Transformer	4 layers, 4 attention heads
Transformer feed-forward dimension	1,024
Policy readout	Classification token
Residual log standard deviation	-1.7 (fixed)
Critic MLP	4 hidden layers, 1,024 units, ReLU
Optimizer	Adam
Actor / critic learning rate	2×10^{-5} / 10^{-4}
Rollout length	32 control steps
Optimization mini-epochs	1
Discount factor / GAE parameter	0.99 / 0.95
PPO clipping threshold	0.2
Gradient-norm limit	50