MaskedManipulator: Versatile Whole-Body Manipulation

Chen Tessler
NVIDIA
Israel

Yifeng Jiang
NVIDIA
USA

Erwin Coumans
NVIDIA
USA

Zhengyi Luo
NVIDIA
USA

Gal Chechik
NVIDIA
Israel

Xue Bin Peng
NVIDIA
Canada
Simon Fraser University
Canada

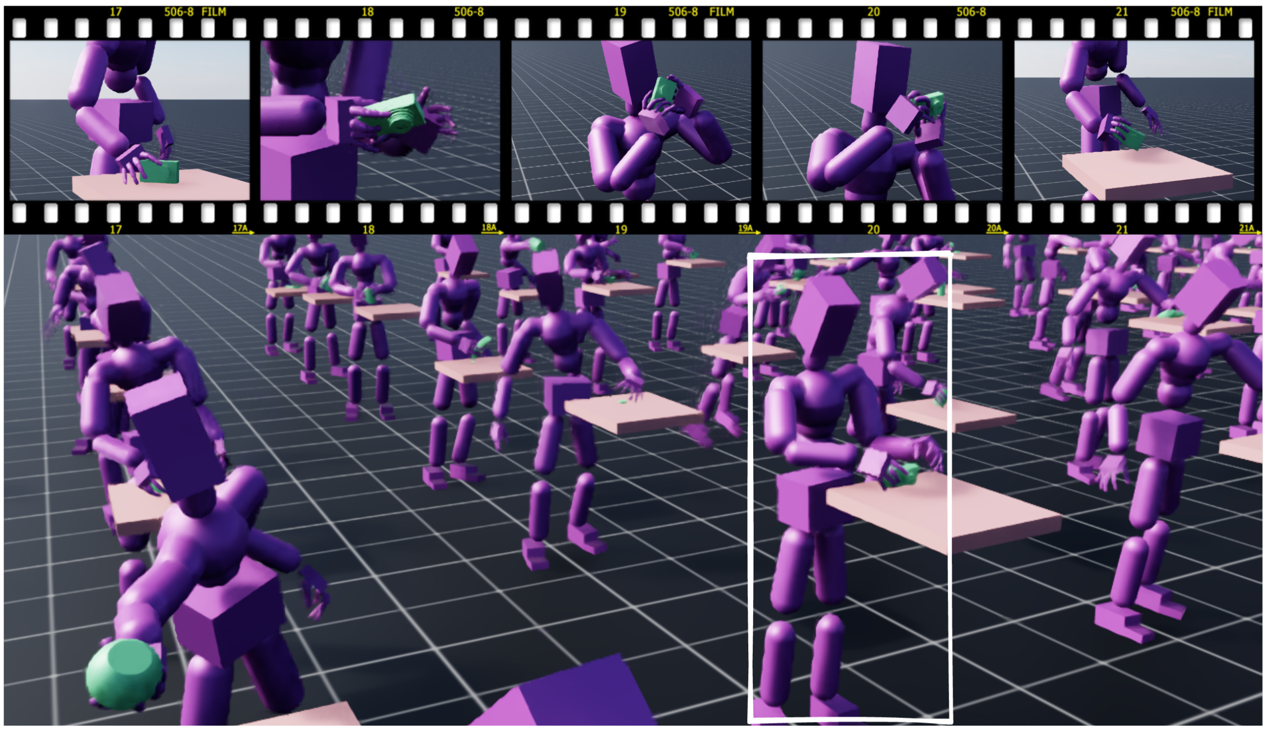


Figure 1: MaskedManipulator enables physics-based humanoids to perform intricate, object interactions from sparse spatio-temporal goals.

Abstract

We tackle the challenges of synthesizing versatile, physically simulated human motions for full-body object manipulation. Unlike prior methods that are focused on detailed motion tracking, trajectory following, or teleoperation, our framework enables users to specify versatile high-level objectives such as target object poses or body poses. To achieve this, we introduce MaskedManipulator, a generative control policy distilled from a tracking controller trained on large-scale human motion capture data. This two-stage learning process allows the system to perform complex interaction behaviors, while providing intuitive user control over both character and object motions. MaskedManipulator produces goal-directed manipulation behaviors that expand the scope of interactive animation systems beyond task-specific solutions.

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763934

CCS Concepts

• **Computing methodologies → Physical simulation**; **Control methods**; **Reinforcement learning**.

ACM Reference Format:
Chen Tessler, Yifeng Jiang, Erwin Coumans, Zhengyi Luo, Gal Chechik, and Xue Bin Peng. 2025. MaskedManipulator: Versatile Whole-Body Manipulation. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers*

'25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3757377.3763934

1 Introduction

Generating versatile humanoid agents capable of autonomous and controllable behaviors is a shared challenge for both character animation and robotics. While recent approaches have significantly advanced locomotion capabilities for simulated [Tessler et al. 2024a; Wu et al. 2025] and real-world humanoids [Allshire et al. 2025; Ji et al. 2024; Ze et al. 2025], synthesizing controllers that can skillfully interact with the environment – especially with small, dynamic objects requiring precise dexterous manipulation – remains a daunting hurdle. A key challenge lies in how to effectively translate high-level human intent into low-level motor commands. For example, walking to an object, picking it up, inspecting the object, and then placing it back down on a table.

We aim to build a versatile controller that seamlessly combines full-body locomotion with dexterous manipulation, which can also interpret such diverse command inputs. However, there exits an inherent tension: the need for broad task flexibility versus the need for exacting physical precision. On one hand, achieving versatility requires controllers capable of understanding abstract user intents, like sparsely defined target coordinates (e.g., "reach this point") or high-level object goals (e.g., "place cup here"). These often underspecified goals present vast solution spaces, posing substantial challenges for control and learning techniques like reinforcement learning (RL). On the other hand, successful manipulation, particularly with common objects, demands meticulous precision. Even small deviations during physical interaction – grasping, repositioning, or placing – can lead to instability, object dropping, or complete task failure, often without recovery [Luo et al. 2024b]. While prior work has tackled aspects of this challenge, existing methods typically excel at either sophisticated manipulation [Iyer et al. 2024; Lin et al. 2024] or versatile body control (e.g., via goal-conditioning [Tessler et al. 2024a]). Achieving unified versatile control that encompasses both body and object goals, and gracefully balances this flexibility-precision trade-off, remains an open problem.

We propose a unified control framework for diverse and precise full-body-manipulation tasks, responding to inputs ranging from detailed kinematic targets (e.g., tracking the hand positions via teleoperation) to sparse, high-level goals (like a desired position of an object). Our framework extends spatio-temporal goal-conditioning (MaskedMimic [Tessler et al. 2024a]) to encompass both the humanoid's body parts and the manipulated object. To master the precise execution required by human-object interactions, our first stage (MimicManipulator) leverages human motion capture (e.g., GRAB dataset [Taheri et al. 2020]). The motion capture data demonstrates successful interaction strategies, such as sequences of grasping, placing, and regrasping objects, and hand-to-hand transfers. This physics-based tracker (MimicManipulator), trained with RL in a fully observable setting, learns the necessary actions to accurately reconstruct these intricate behaviors. Then, to enable versatile control from sparse goals, MimicManipulator's expert interaction knowledge is distilled into our MaskedManipulator policy. By learning from MimicManipulator, MaskedManipulator becomes capable of producing human-like and diverse behaviors in response to under-specified goals for both body parts and objects (e.g., "bring the object to coordinate (x, y, z)"), effectively addressing the multi-solution challenge of versatile control.

The central contributions of this work are:

(1) We extend the versatile full-body controller, MaskedMimic [Tessler et al. 2024a], enabling full-body-manipulation. We demonstrate how to successfully leverage human demonstrations to produce diverse, physically plausible, and human-like solutions to underspecified tasks involving coupled human-object interactions.
(2) MimicManipulator: A physics-based motion tracker that accurately infers actions and reconstructs dexterous human full-body-manipulation sequences from reference motion data.
(3) MaskedManipulator: A unified generative full-body-manipulation policy that enables diverse spatio-temporal goal-conditioning for both humanoid body parts and manipulated objects.

2 Related Work

Our work builds upon advances in demonstration-driven control, object manipulation, and versatile goal-conditioned policies for physics-based characters.

Demonstration-driven control. Traditional reinforcement learning (RL) for character control often requires meticulous reward engineering. Imitation learning (IL) offers a powerful alternative by leveraging expert demonstrations. DeepMimic [Peng et al. 2018] showed RL can robustly replicate reference motions, a paradigm extended to object manipulation by works like [Yu et al. 2025] for dexterous sequences and InterMimic [Xu et al. 2025] for retargeting interactions across morphologies. Our MimicManipulator stage is inspired by these approaches, using RL to train a single policy to precisely track the entire diverse human full-body-manipulation demonstrations from GRAB [Taheri et al. 2020], thereby recovering actions and ensuring physical plausibility.

Object manipulation and grasping. Integrating object manipulation into humanoid control substantially increases task complexity. Early approaches achieved progress without reference data by relying on hand-crafted rewards and large-scale reinforcement learning [Akkaya et al. 2019; Lin et al. 2024]. More recent work has explored demonstration-driven methods: for example, OmniGrasp [Luo et al. 2024a] synthesizes full-body motions conditioned on dense object trajectories and trained via dense rewards. It achieves this by leveraging a scene-agnostic generative prior that is steered through hierarchical reinforcement learning. Compared to our approach, this design presents two key limitations: (1) the reliance on dense reward signals restricts OmniGrasp to short-horizon training objectives, making it unsuitable for more general goal-conditioned tasks such as bringing an object to a target position [Pignatelli et al. 2023]; (2) the scene-agnostic prior hinders dexterous control. As a result, OmniGrasp struggles reproducing dexterous behaviors such as grasping a teapot by its handle or performing hand-to-hand transfers. In contrast, our method first learns to reproduce dexterous human-object demonstrations from dense goals and then distills this knowledge into a flexible, long-horizon, goal-conditioned controller. The resulting controller, MaskedManipulator, supports a range of object and/or humanoid body-part goals, enabling the synthesis of diverse full-body-manipulation behaviors from sparse constraints and supporting more adaptable, long temporal sequence control.

Versatile and goal-conditioned control. A trend in robotics and animation involves hierarchical control, where high-level systems specify abstract goals (e.g., walk targets [Peng et al. 2022], grasp poses [Li et al. 2024], 2D paths [Li et al. 2025]) for low-level execution, akin to System 1/System 2 reasoning [Kahneman 2011]. MaskedMimic [Tessler et al. 2024a] significantly advanced System1 versatility by framing character control as motion inpainting, allowing a unified model to synthesize diverse motions from sparse future joint targets. Our work extends this by enabling explicit spatio-temporal goal-conditioning on both humanoid body parts and the manipulated object within MaskedManipulator. This provides direct, versatile control over the entire coupled human-object system, crucial for sophisticated full-body-manipulation tasks.

3 Preliminaries

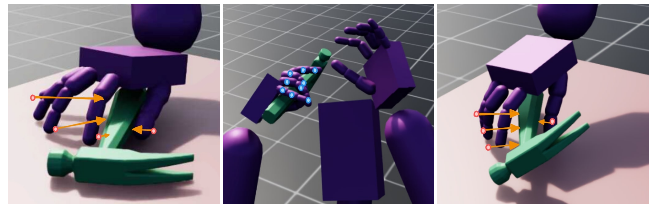
We formulate our problem within the framework of goal-conditioned Markov Decision Process (MDP) [Puterman 2014], defined by a tuple $M = (S, G, A, P, R, \gamma)$, consisting of states S, goals G, actions A, transition dynamics P, a reward function R, and a discount factor γ. The objective is to learn a policy $\pi(a_t|s_t, g_t)$ that selects an action a_t based on the current state s_t and a task-specific goal g_t.

The state of the humanoid's J links at time t is described by their individual 3D positions $p_{j,t} \in \mathbb{R}^3$ orientations $\theta_{j,t}$, linear $v_{j,t}$ and angular $\omega_{j,t}$ velocities. Each orientation $\theta_{j,t}$ is represented using a continuous 6D rotation vector [Zhou et al. 2019]. All link positions and orientations constitute the full body kinematic state, denoted $q_t = (\{p_{j,t}\}_{j=1}^J, \{\theta_{j,t}\}_{j=1}^J)$. The corresponding linear and angular velocities collectively form $\dot{q}_t$. When interacting with the scene, a body part may be subjected to contact forces, which we denote by $c_{j,t}$. Object states q_t^{obj} are similarly defined by 3D position p_t^{obj}, 6D orientation θ_t^{obj}, velocities $\dot{q}_t^{\text{obj}}$, and contact forces c_t^{obj}.

Reference kinematic and contact quantities, sourced from Motion Capture (MoCap) data or a target trajectory generator, are denoted with a hat accent (e.g., $\hat{q}_t, \hat{\dot{q}}_t, \hat{q}_t^{\text{obj}}, \hat{c}_t, \hat{c}_t^{\text{obj}}$). In contrast to the simulated information, the reference $\hat{c}_t$ are contact indicators. Quantities without this accent refer to states within our physics simulation.

4 Method

Our approach for generating versatile humanoid full-body-manipulation capabilities is structured around a two-stage learning process. Human motion capture data provides rich kinematic descriptions of complex interactions but typically lacks the underlying low-level motor commands (actions) required to drive a physics-based character. Our first stage, MimicManipulator, addresses these challenges by training a high-fidelity motion tracker within a physics-based simulation. Operating in a full-information setting with respect to the reference kinematics, MimicManipulator learns not only to recover the necessary actions but also to physically reconstruct the target full-body-manipulation sequences, thereby ensuring dynamic feasibility and emphasizing the nuances of object handling.



(a) Pre-contact **(b) During contact** **(c) Post-contact**

Figure 2: Phased contact reward for precise manipulation. We design a three-stage contact reward: (a) Approach: Tracks the reference motion while aligning the hand's path relative to the object surface. (b) Engagement: Ensures critical contacts are maintained according to the reference. (c) Release: Promotes a smooth and timely object disengagement mirroring the demonstration. The spheres illustrate the reference joints and the arrows illustrate vectors to the closest point on the object's surface.

Subsequently, the robust interaction skills learned by MimicManipulator is distilled into a versatile control policy, MaskedManipulator. This second stage yields a policy that can generate novel, physically plausible behaviors by dynamically conditioning on spatio-temporal goals specified for various parts of the humanoid body or the manipulated object itself. The following sections detail the objectives, architecture, and training procedures for each of the two stages.

4.1 MimicManipulator

The first stage focuses on reproducing the intricate behaviors required for dexterous full-body-manipulation. We train a single policy π_{track}, a motion-tracker denoted by MimicManipulator, to accurately reproduce the complex full-body-manipulation sequences from the processed GRAB dataset. The reference kinematic data provides a rich description of the desired high-level behavior but does not contain the low-level motor commands (actions a_t) necessary to execute these motions in a physically simulated environment. Therefore, we formulate this task as a reinforcement learning problem where π_{track} learns to infer these actions by imitating the reference motion.

4.1.1 Task design.

Observation. To enable object interactions, the observation s_t comprises of the character's proprioception q_t, object information q_t^{obj} and character-object relational features. All components are expressed in a character's local coordinate frame and aligned with it's current root heading (yaw) for policy invariance. The character-object relational features capture the spatial relationship between the humanoid and the object, by providing the vector (direction and distance) from each body part to the object surface. **Object structure:** Precise interaction requires awareness of the object geometry. To accomodate this, we also provide a Basis-Point-Set [Prokudin et al. 2019, BPS] representation, which represents all objects by the surface distances to a pre-sampled set of points. Compared to point-cloud representations, the pre-sampled and fixed BPS allow processing using simpler architectures (e.g., MLP), allowing faster training and inference, while still being expressive. **Future reference trajectory (goal)**: To reproduce the target motions, MimicManipulator observes the the goal future poses in the next frame $K = 1$ and $K = 30$ frames (1 second), denoted by g_t^{track}. Each target future pose observation $\{\hat{q}_{t+k} \Theta q_t\} = \{\hat{p}_{t+k} - p_t, \hat{q}_{t+k} \Theta q_t\}$ is a future pose $\hat{q}_{t+k}$ normalized with respect to the current simulated pose q_t, where Θ represents the inverse quaternion. In addition, we provide $\hat{c}_{t+k}$ to indicate which body parts should have contact with the object.

Reward and termination. Successfully reconstructing long and intricate human-object interaction sequences requires high precision. Minor inaccuracies during critical interaction phases (e.g., grasping, placing) can compound, leading to failures much later in the motion. This delayed consequence creates a challenging credit assignment problem for reinforcement learning.

Reference demonstrations implicitly encode successful, precise interaction strategies. For example, where to grab the object in order to naturally reach a target orientation, or to successfully perform a hand-to-hand transfer. We leverage this data and design reward R_{track} and termination conditions to synergistically guide the MimicManipulator agent towards accurate and physically plausible reconstructions.

The reward R_{track} is a product of several components. We found that the multiplicative design helps learning precise, full-body-manipulation over long temporal sequences by strictly requiring competence in all elements of the complex motion [Park et al. 2019; Won et al. 2020; Xu et al. 2025].

$$R_{\text{track}} = r^{\text{pose}} \cdot r^{\text{contact}} \cdot r^{\text{energy}} \cdot r^{\text{interaction}} . \tag{1}$$

Each component $r^{(\cdot)}$ is an exponential of a negatively weighted error or cost (e.g., $r = \exp(-w \cdot \text{cost})$), valued between (0, 1].

$$r^{\text{humanoid}} = r^{\text{ht}} \cdot r^{\text{hr}} \text{ and } r^{\text{obj}} = r^{\text{ot}} \cdot r^{\text{or}} \tag{2}$$

components consider the translation and rotation errors for the human and object respectively. The energy component attempts to mitigate artifacts such as jitters. The interaction component aims to prevent human-object penetrations by encouraging a gentle approach and interaction.

The contact component, illustrated in Figure 2, is formulated to encourage the policy to reproduce the contact patterns and timing observed in the reference motion. The reward is structured into three distinct phases, each corresponding to a specific stage of the interaction, thereby guiding the agent to accurately replicate the demonstrated behavior. To prevent abrupt behaviors, the approach and engagement phase starts 1 second before contact and ends 0.2 seconds after the reference motion indicates contact has started. Similarly, the disengagement phase beins 1 second before the reference motion indicates that contact has ended. We provide additional details in the supplementary material.

In the **approach and engagement** phase, we identify future contact bodies from the reference trajectory. For each future contact body, the closest point on the object mesh is determined in both the simulation and the reference motion. The reward then penalizes discrepancies in distances and surface normals between each simulated and reference contact body, thereby promoting similarity in the approach trajectory and orientation. This mechanism ensures that end-effectors (e.g., hands) are guided smoothly toward the target contact regions in a manner consistent with the demonstration.

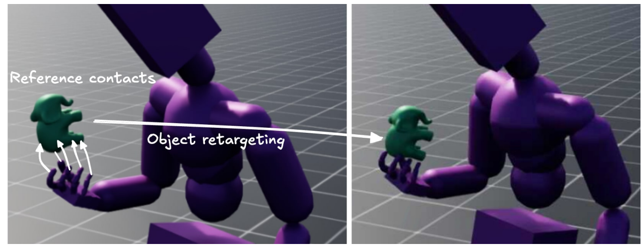


Figure 3: Object Retargeting for Morphological Differences. Transferring motion between characters of varying shapes can misalign human-object interactions (left). Our method leverages original contact data to retarget the object's trajectory, preserving interaction consistency (right).

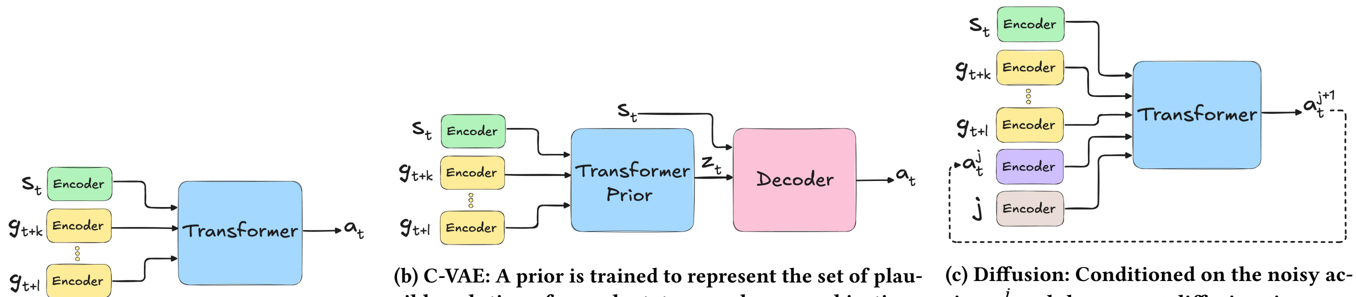
During the **contact maintenance** phase, we identify which reference body parts $\hat{c}_{j,t}$ are marked as should-be-in-contact. For each such body part, a reward is computed based on its proximity to the object mesh. This formulation encourages stable contact, discourages premature detachment, and ensures alignment of the contact configuration with that of the reference motion.

Finally, in the **disengagement phase**, previously contacting body parts are incentivized to move away from the object in a controlled and timely manner. Rewards are again based on distance and surface normal alignment, encouraging a smooth release that mirrors the reference disengagement dynamics.

Complementing the reward design, we leverage early termination to define a feasibility envelope for the training phase [Luo et al. 2023; Peng et al. 2018; Tessler et al. 2024a]. An episode is terminated if any body part deviates by >25cm from reference, or the object deviates >10cm from reference, or when there's a prolonged unintended loss of contact for >10 consecutive frames during a contact phase, or if the character maintains contact >0.4sec after the reference has released.

Prioritized training for complex interactions. We train a single policy to reconstruct the entire dataset. The reference sequences span a wide range of difficulty. This variation arises from both the object's characteristics (e.g., grasping a mug by its handle is more demanding than grasping a banana) and the intricacy of the interaction itself (e.g., bimanual operations or tasks requiring precise tool-like dexterity). During training we periodically evaluate the agent's performance and over-sample sequences in which it has failed [Luo et al. 2024a, 2023; Tessler et al. 2024a].

4.1.2 Data processing. The GRAB dataset consists of motion sequences collected across 10 different human subjects, differing by gender and size. This poses a challenge for training the controller, as it requires awareness and generalization across different morphologies. To simplfy the control problem, we use a single, canonical humanoid body shape (mean SMPL-X). The core difficulty lies in accurately retargeting the diverse source motions from different subjects to our canonical character while preserving the semantically salient characteristics of the original human-object interaction.



(a) Deterministic

(b) C-VAE: A prior is trained to represent the set of plausible solutions for each state s_t and sparse objectives $g_t^{\text{versatile}}$.

(c) Diffusion: Conditioned on the noisy action a_t^j and the current diffusion timestep j.

Figure 4: MaskedManipulator architectures: We illustrate the inference procedure of the considered architectures.

We propose a simple yet effective approach. First, we map the motion from source to target skeletons by transferring the local DoF rotations. Transferring rotations to a new body size leads to misaligned human-object motion. For example, when rotated downwards, a longer upperarm will result in the hand being beneath the object. Our second step is object retargeting. We extract contacts from the human-object aligned source motion. This tells us which body part is in contact with the object's mesh, and where. We then retarget the object's global position to best preserve these contact relationships with the canonical character's end-effectors (additional details in the supplementary material and illustrated in Figure 3).

Despite retargeting, some motions remain unsuitable for physical reconstruction. We filter these, primarily due to: Complex bimanual interactions where our object-centric retargeting may not fully resolve hand-object-hand alignment (e.g., transferring motion to a shorter character leads to a large gap between hands), and interactions reliant on features our simplified humanoid model lacks (e.g., placing sunglasses on a detailed face). This data processing, including retargeting and filtering, yields a training set of 1007 motion sequences and a test set of 141 sequences (GRAB subject 10).

4.2 MaskedManipulator

The second stage of our framework distills the rich, physically-grounded interaction expertise learned by π_{track} (from MimicManipulator) into a versatile, generative control policy, $\pi_{\text{versatile}}$ (our MaskedManipulator). The ultimate goal for MaskedManipulator is to produce diverse, human-like, and physically plausible full-body manipulation behaviors in response to sparse future goals. These goals might specify the desired pose for a hand, the target location for an object, or a full-body stance at a future time, effectively leaving many other aspects of the motion for the policy to "inpaint".

Directly training such a versatile policy from scratch is exceptionally challenging. Sparse goals often lead to a misspecified objective with an overwhelmingly large space of potential solutions, posing significant exploration and credit assignment difficulties. However, MimicManipulator (π_{track}), having learned to reconstruct a wide array of complex human demonstrations, implicitly captures a distribution of high-quality, physically feasible solutions to various interaction sub-tasks. Our approach extends the methodology of MaskedMimic [Tessler et al. 2024a], which demonstrated that framing control as a "physics-based inpainting" problem – where the policy fills in missing parts of a motion sequence – can effectively learn versatile behaviors from diverse data.

Goal Specification and Distillation: The core idea is to train MaskedManipulator ($\pi_{\text{versatile}}$) to predict the detailed actions that MimicManipulator (π_{track}) would execute to complete a motion, but conditioned only on the current state s_t and a sparse set of future goals $g_t^{\text{versatile}}$. We employ an online teacher-student distillation process based on DAgger [Ross et al. 2011].

At each step, the student policy $\pi_{\text{versatile}}$ receives the current state s_t and a randomly masked version of the future reference trajectory, $g_t^{\text{versatile}}$. This sparse goal $g_t^{\text{versatile}}$ specifies target future states (e.g., pose of a hand, object location, or full body stance) for a subset of entities at one or more future timesteps. All other future state information is masked. The student then predicts an action $a_t^{\text{versatile}}$ which is used to control the character in the simulation.

Concurrently, the teacher policy π_{track} observes the same current state s_t but with the unmasked, dense future reference trajectory g_t^{track} (as used during its own training). The distillation objective is to minimize the L2 distance between the actions predicted by the student and the teacher:

$$\mathcal{L}_{\text{distill}} = -\log \pi_{\text{versatile}} \left(a_t^{\text{track}} | s_t, g_t^{\text{versatile}} \right) \qquad (3)$$

We adapt the structured masking scheme from MaskedMimic [Tessler et al. 2024a] to define $g_t^{\text{versatile}}$, crucially extending it to include conditioning on the future pose of the manipulated object alongside humanoid body parts.

Policy Architecture and Training: To accommodate a variable number of sparsely specified future goal constraints, each defined by a target pose and a future timestep, MaskedManipulator utilizes a transformer architecture [Vaswani et al. 2017]. The input to the transformer is a sequence of tokens. Fixed-information inputs, such as current proprioception (joint states, root state), hand-to-object vectors, object and table pose, and object and table BPS, are jointly mapped to a shared token. Each specified future goal (e.g., target pose of hand j at time $t + k$) is also encoded into its own token. Which each target future pose is represented using a unique token, they are generated using a shared encoder.

Given that a sparse goal $g_t^{\text{versatile}}$ can often be achieved through multiple valid and human-like motion sequences (i.e., the problem is multi-modal), we experiment with three classes of policy architectures for $\pi_{\text{versatile}}$, illustrated in Figure 4:

(1) Deterministic: This standard approach learns to predict a single, mean action based on the teacher's demonstrations. While computationally efficient, this can average over the diverse solutions present in the data, potentially leading to less expressive or less performant behaviors when multiple valid options exist.

(2) Conditional Variational Autoencoder (C-VAE): To explicitly model the multi-modality of solutions, we employ a C-VAE architecture with a learned prior, drawing inspiration from [Rempe et al. 2021; Tessler et al. 2024a]. The prior network takes the sparse goal $g_t^{\text{versatile}}$ and current state s_t to predict a latent distribution $\mathcal{N}(\mu_{prior}, \sigma_{prior})$ representing plausible general solutions. During training, a residual encoder provides the offset in the latent space, predicting the precise solution from the reference data. We extend the architecture in MaskedMimic. There, the encoder observes both the reference motion g_t^{track} and the sparse goals $g_t^{\text{versatile}}$. As a result, the encoder is required to implicitly predict both the prior distribution and the offset. We simplify this design by providing the encoder with the prior's output $\mathcal{N}(\mu_{prior}, \sigma_{prior})$. We found it reduces parameter count and improves performance. During inference, sampling from the learned prior allows MaskedManipulator to generate diverse actions satisfying the sparse goal.

(3) Diffusion Policy: As an additional generative approach, we investigate a diffusion-based policy [Chi et al. 2023; Ho et al. 2020; Lu et al. 2025]. Unlike C-VAEs, diffusion models do not require an explicit prior network or separate encoder for the posterior during training for their generative process. The policy is trained to denoise an action that is initially pure Gaussian noise, iteratively refining it over N steps to produce a clean action a_t^N. This iterative refinement process is conditioned on the current state s_t and sparse goal $g_t^{\text{versatile}}$. To provide the necessary conditioning for each denoising step j, we augment our transformer architecture with two additional input heads: one tokenizing the noisy action a_t^j from the current step, and another representing the current denoising timestep j. At each step j, the policy predicts the less noisy action a_t^{j-1} (or equivalently, the noise to remove). The final denoised action a_t^0 (conventionally, or a_t^N if N is the start of denoising from pure noise) is used to control the character. In addition to Diffusion using DAgger, we also experiment with a fully-offline version, where MaskedManipulator is trained using supervised learning on trajectories collected by MimicManipulator.

Both the C-VAE and Diffusion approaches are explicitly chosen for their capacity to model the inherent diversity and multi-modality in the solution space when presented with sparse goals. This empowers MaskedManipulator not merely to fulfill specified constraints, but to do so with a rich variety of human-like behaviors, reflecting the breadth of solutions learned by MimicManipulator.

5 Experiments

Experimental setup: All experiments are conducted using the ProtoMotions framework [Tessler et al. 2024b] with IsaacLab [Mittal et al. 2023] as the underlying physics simulator. The simulation runs at 120 FPS, and with 4 physics decimation steps, our control policies operate at an effective rate of 30 FPS. To manage interpenetrations robustly, we set the simulation's depenetration velocity to 100. To better approximate the contact dynamics of human soft tissue interacting with objects using rigid body simulation, we use a global friction coefficient of 1.5.

Training details: All MimicManipulator tracking policies (π_{track}) are trained using Proximal Policy Optimization (PPO) [Schulman et al. 2017]. The MaskedManipulator variants ($\pi_{\text{versatile}}$) are trained using DAgger [Ross et al. 2011] with π_{track} as the expert teacher. Each policy variant is trained for approximately 40,000 episodes, distributed across 8 NVIDIA A100 GPUs. Specific hyperparameters for PPO and the distillation process are detailed in the Appendix.

Evaluation metrics: The full-body-manipulation sequences from GRAB are often long and complex, involving multiple pick-and-place iterations, and hand-to-hand transfers. To comprehensively evaluate performance, we employ several metrics:

Full-Sequence Success Rate: The percentage of test sequences successfully completed from start to finish.

First-Interaction Success Rate: To isolate early interaction performance from later complexities, this measures success from the motion's start until the second required contact with the object.

Mean Per Joint Position Error (MPJPE): The average Euclidean distance between the simulated joint positions and the reference motion's joint positions, providing a measure of tracking fidelity.

Average Sequence Length: The average episode length until termination is encountered (or episode ends).

We define failure (leading to episode termination for success rate calculation) if any humanoid joint deviates by more than 50cm from its reference position, or if the manipulated object deviates by more than 25cm from its reference position [Luo et al. 2024a].

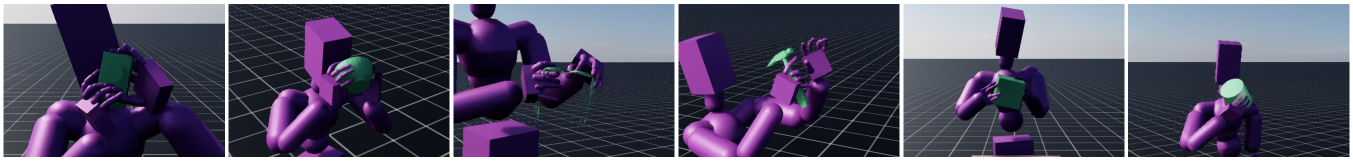
6 Results

6.1 MimicManipulator

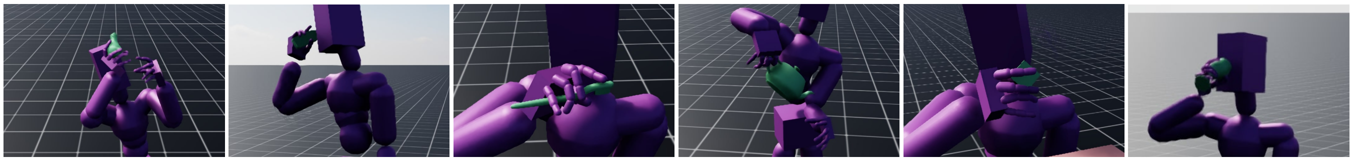
We evaluate ManipulationMimic's performance through ablation studies and quantitative comparisons, and provide qualitative examples of its capabilities.

Ablation study and quantitative analysis. Table 1 presents MimicManipulator's performance on the GRAB dataset and the impact of key design choices. We ablated: (1) tight early termination criteria, (2) prioritized training on harder motions, and (3) the phased contact guidance reward. The results demonstrate that each component positively contributes to the overall tracking success, with the strict early termination yielding the most significant improvement by maintaining a tight feasibility envelope.

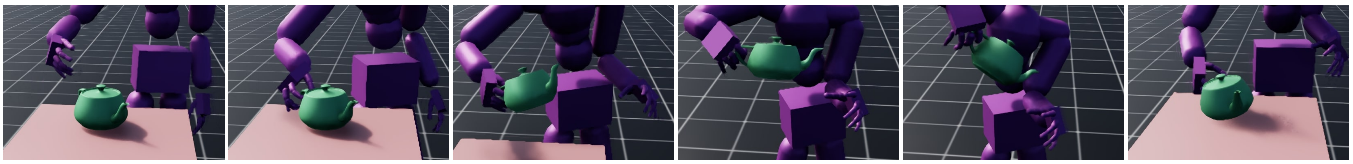
We also compare MimicManipulator with our implementation of InterMimic [Xu et al. 2025]. For a fair comparison, we trained a single InterMimic model on the entire GRAB training set, similar to MimicManipulator, omitting InterMimic's subject-specific policy distillation stages. In addition, we did not include physical state initialization as it introduced instabilities in dexterous manipulation tasks. MimicManipulator outperforms this adapted InterMimic baseline. We attribute InterMimic's lower performance in this setting to its potentially looser tracking requirements and a lack of explicit object structure awareness (e.g., a BPS representation for



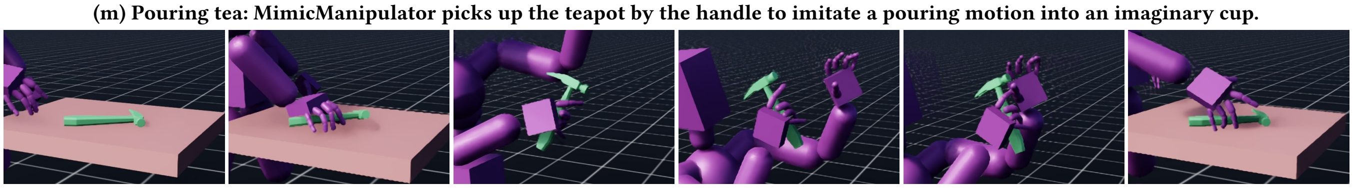
(a) Taking a picture (b) Drinking soup (c) Cleaning glasses (d) Using a hammer (e) Inspecting a cube (f) Inspecting object



(g) Screwing lightbulb (h) Using a phone (i) Brushing teeth (j) Pouring tea (k) Inspecting a cube (l) Drinking wine



(m) Pouring tea: MimicManipulator picks up the teapot by the handle to imitate a pouring motion into an imaginary cup.



(n) Using a hammer: MimicManipulator precisely handles a hammer, maintaining proper orientation while demonstrating how to hit a nail.

Figure 5: MimicManipulator – full-body tracking: MimicManipulator successfully interacts with a wide range of objects, reconstructing various daily human behaviors. ***Textual descriptions are provided for the reader's convenience. The model itself is not conditioned on textual labels.***

Table 1: MimicManipulator: We ablate various design decisions, stripping another component and measuring the cumulative importance. E.g., the "Contact guidance" is without "Tight termination" and "Prioritized scenes". We also compare with our implementation of InterMimic [Xu et al. 2025] (see Section 6.1).

	Train				Test			
	Sequence Success	First Sequence Success	MPJPE [mm]	Tracking Length [s]	Sequence Success	First Sequence Success	MPJPE [mm]	Tracking Length [s]
MimicManipulator (ours)	80.7%	93.5%	9.8	9.6	60.2%	83.7%	13.2	7.3
(-) Tight termination	74.6%	91.2%	15.3	9.3	51.8%	80.1%	22.8	7
(-) Prioritized scenes	71.2%	89.3%	15.5	9.2	56%	73.8%	18.4	7
(-) Contact guidance	69.4%	86.6%	16.4	9.2	47.5%	78%	25.5	6.5
InterMimic* [Xu et al. 2025]	11%	31.1%	42.2	6.2	8.5%	29.8%	50.5	4.9

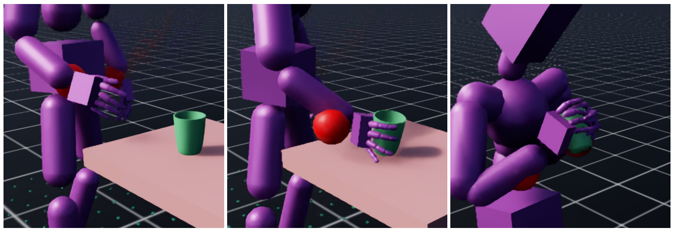
contact guidance) which is crucial for precise interactions with diverse objects in GRAB.

Qualitative analysis. As illustrated in Figure 5 and the supplementary video, MimicManipulator successfully reconstructs a diverse range of complex, full-body interactions while preserving natural, human-like motion qualities. For instance, the policy can accurately simulate grasping a teapot by its handle to pour liquid, or picking up a hammer and striking a nail. These examples highlight its ability to learn nuanced, tool-specific behaviors. While MimicManipulator reliably reconstructs the majority of motions, occasional uncanny behaviors can be observed, particularly during the precise moments of contact initiation or release. These may stem from the discrete nature of the termination conditions related to contact, which can sometimes force abrupt transitions rather than perfectly smooth engagement or disengagement.

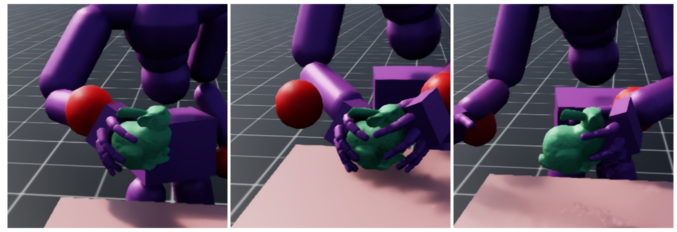
6.2 MaskedManipulator

We quantitatively evaluate MaskedManipulator ($\pi_{\text{versatile}}$), assessing its versatility, generalization, and long-horizon reasoning.

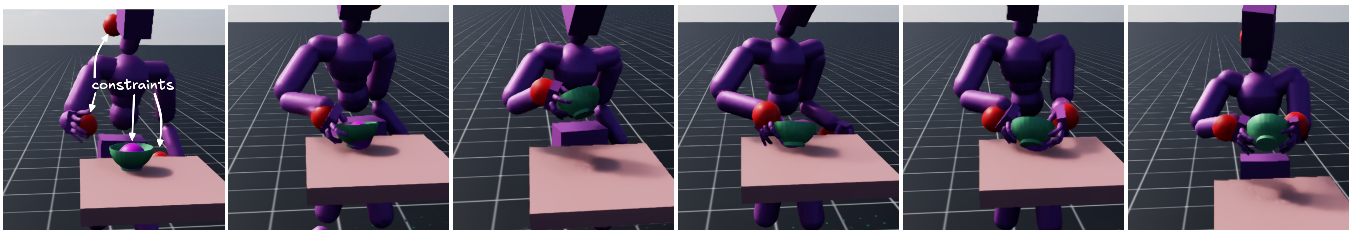
Quantitative analysis. **Teleoperation-Style Pose Matching:** Table 2 compares the proposed MaskedManipulator architectures (Deterministic, C-VAE, Diffusion) on a teleoperation-like task. Here,



(a) **Wrist control: Conditioned only on the wrist position and orientation, the agent predicts when the motion is intended to interact with an object.**



(b) **Wrist control, hand-off: When only conditioned on the wrists, the agent is required to 'guess' which hand is more likely to maintain contact.**



(c) **Simulating teleoperation with object control: Conditioning on the head position provides additional information on the body position, while conditioning on the object pose helps the policy infer how the hands should interact with the object.**

Figure 6: MaskedManipulator – wrists, head, and/or object conditioning: MaskedManipulator is conditioned on when and where to transport the object. The red spheres indicate target joint positions, whereas the purple where indicates the target object position.

the policy is conditioned on near-future goals for head and wrist poses, plus the object's pose, aiming to accurately achieve these combined configurations. While training performance was similar for the stochastic C-VAE and Diffusion models, the Diffusion policy generalized better, successfully completing 3.6% more unseen test sequences than the other two. This highlights the advantage of its capacity to model complex, multi-modal action distributions for robust generalization. In addition, we observe a dramatic performance drop when training on offline data, highlighting the importance of self-play for learning to recover from mistakes [Ross et al. 2011].

Long-Horizon Sparse Object Goal Chaining: We further test the policies on a challenging long-horizon task (Table 3) where MaskedManipulator receives a sparse object pose goal 2 seconds into the future. Upon reaching the target time, a new 2-second future object goal is set, repeating until the reference motion ends. This evaluates the policy's ability to reason about distant object states while producing appropriate full-body motion and stable grasps for diverse objects. Performance is measured by the success rate in achieving the sequence of object goals. Consistent with the teleoperation task and prior work [Tessler et al. 2024a], modeling solution diversity proved crucial. While the C-VAE performed better on the training set (indicating some overfitting), the Diffusion model again exhibited superior generalization, successfully solving 9.3% more unseen test sequences than the C-VAE.

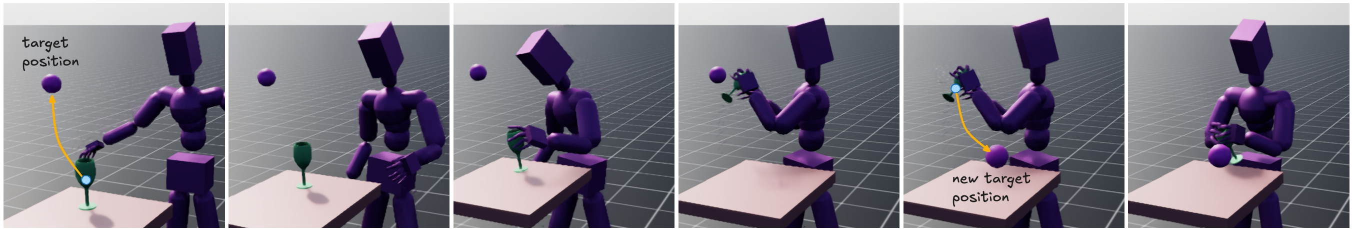
Qualitative analysis. In Figures 6 to 8, we qualitatively demonstrate MaskedManipulator's versatility and the nuanced behaviors it can produce. As shown in Figure 6, when given goals for wrist position and rotation, MaskedManipulator achieves stable and accurate pose tracking. Interestingly, even without explicit instructions, the policy often infers when to make or break contact with objects to facilitate the desired poses, highlighting its learned understanding of physical interaction. In Figure 7, we show that MaskedManipulator can accomplish sparse object goals specified several seconds into the future. For example, transferring an object to a target location by generating the necessary full-body motion and even hand-to-hand object transfer when needed. Finally, Figure 8 highlights purely

Table 2: 'Teleoperation': We compare the various architectures on a teleoperation task. Here, MaskedManipulator is conditioned on the positions and rotations of the head, wrists, and object.

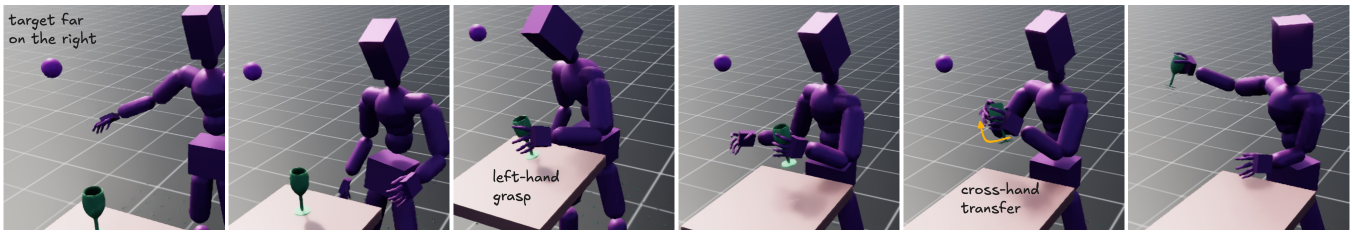
	Train		Test	
	Success	MPJPE [mm]	Success	MPJPE [mm]
Deterministic	73.8%	16.4	54.6%	24
C-VAE	78.5%	12.7	54.6%	24.4
Offline Diffusion	50%	27.8	25.5%	38.5
Diffusion	78.6%	12.1	58.2%	19.7

Table 3: Object goals: The agent is provided a sparse objective indicating where (and when) it should transport the object to.

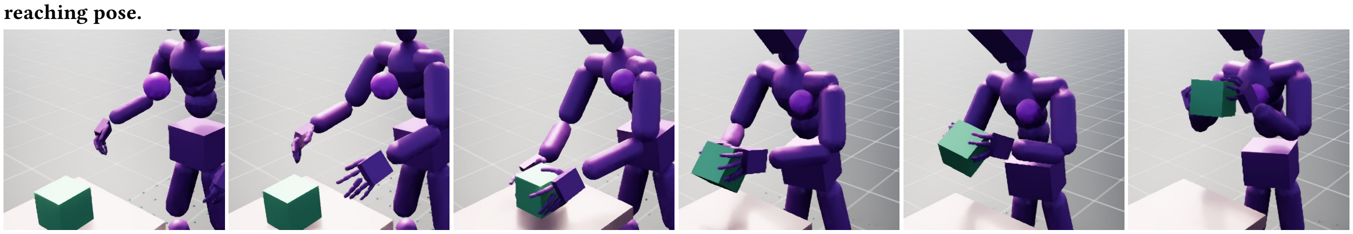
	Train		Test	
	Success	Tracking Length [s]	Success	Tracking Length [s]
Deterministic	54.9%	7.8	46.1%	6.2
C-VAE	64.7%	8.3	50.3%	6.6
Offline Diffusion	30.9%	6.8	32.6%	5.8
Diffusion	57.3%	7.9	59.6%	6.8



(a) Sequential goals: The agent picks up and transports the object to the target position. It reacts to changes in the objective.



(b) Solution adaptation: The agent picks up the object with its left hand. It then transfers the object to the right hand for a more natural reaching pose.



(c) Bi-manual: MaskedManipulator uses both hands when tasked with lifting a large object.

Figure 7: MaskedManipulator – full-body motion from sparse object goals: MaskedManipulator is conditioned on when and where to transport the object.

generative behavior when the policy runs with minimal or no explicit goals. In such cases, MaskedManipulator reproduces natural interaction patterns from the GRAB training distribution, such as picking up a toy airplane and "flying" it through the air, showcasing its ability to generate contextually relevant, human-like actions without direct prompting.

These qualitative examples underscore MaskedManipulator's ability to not only follow specified goals but also to generate rich, physically grounded, and contextually appropriate behaviors learned from the human demonstration data.

7 Limitations

While our framework demonstrates significant progress in versatile full-body-manipulation, we identify several limitations that offer opportunities for future research:

Fidelity of the Unified Tracker (MimicManipulator): We observed that training MimicManipulator to track the full diversity of the GRAB dataset, while successful, sometimes results in slightly lower reconstruction fidelity for individual complex sequences compared to a tracker specialized (i.e., overfit) to only that single motion. This suggests that the current policy representation or capacity might be a bottleneck when learning a very broad repertoire of intricate skills. Future work could explore more expressive policy architectures or adaptive mechanisms within the tracker.

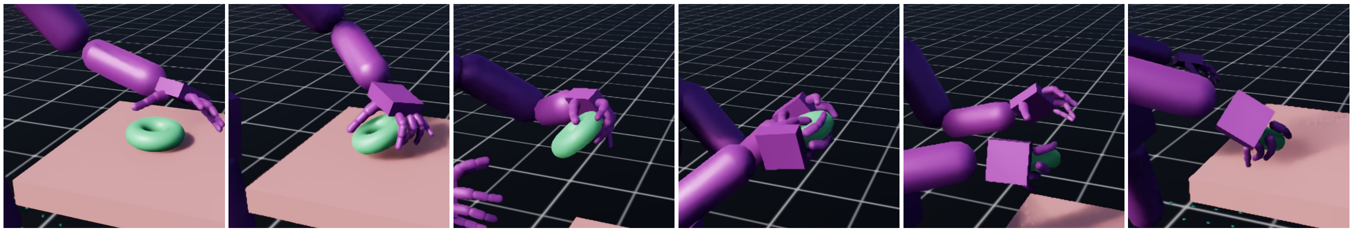
Reconstruction Coverage: Despite our efforts in data processing and robust reward design, MimicManipulator does not reconstruct all motions from the processed GRAB dataset. Certain interactions remain challenging for the current system to reproduce faithfully within the physics simulation, and the performance of MaskedManipulator on new and unseen tasks leaves room for improvement.

Granularity of Control in MaskedManipulator: Currently, MaskedManipulator infers a significant amount of the interaction behavior (e.g., specific grasp points on an object, precise timing of contact engagement) based on the sparse goals and its learned priors from MimicManipulator. It does not allow for explicit user control over how an interaction should be performed at a fine-grained level, such as specifying exact contact locations on the object surface.

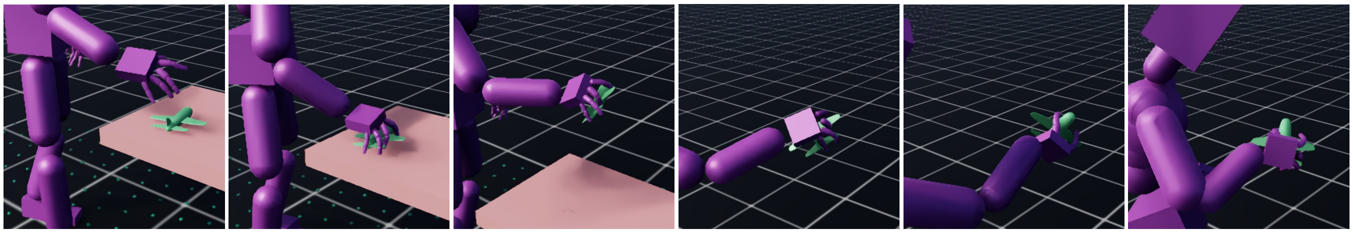
Addressing these limitations could lead to more capable, precise, and controllable virtual humans, further bridging the gap between high-level creative intent and realistic, interactive animation.

8 Conclusions

In this work, we presented a novel two-stage framework for versatile, physics-based full-body-manipulation, enabling digital humanoids to perform complex interactions with objects in response to intuitive, high-level goals. Our approach bridges the gap between the need for precise physical execution and the desire for flexible, versatile control. The resulting MaskedManipulator successfully



(a) Inspecting a large torus: The agent picks up the object with its left hand, holds it with both hands while inspecting, transfers to the right hand, and puts it down.



(b) Flying an airplane: Observing only the object position and BPS [Prokudin et al. 2019], the agent "flys" the toy airplane through the air.

Figure 8: MaskedManipulator – generative interaction: When only conditioned on the pelvis position, MaskedManipulator generates novel interactions.

synthesizes diverse and physically plausible full-body behaviors from sparse, underspecified goals, effectively navigating the multi-modal solution spaces inherent in versatile control.

Our experiments demonstrate that this framework can produce sophisticated manipulation behaviors, such as precise object placements, hand-to-hand transfers, and dynamic interaction with tools, all driven by simple goal specifications. The ability to control both the character and the object through a unified interface opens up exciting possibilities for animators and creators of interactive virtual experiences, simplifying the creation of nuanced and believable character actions.

References

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. 2019. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113* (2019).

Arthur Allshire, Hongsuk Choi, Junyi Zhang, David McAllister, Anthony Zhang, Chung Min Kim, Trevor Darrell, Pieter Abbeel, Jitendra Malik, and Angjoo Kanazawa. 2025. Visual Imitation Enables Contextual Humanoid Control. *arXiv preprint arXiv:2505.03729* (2025).

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. 2023. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* (2023), 02783649241273668.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Aadhithya Iyer, Zhuoran Peng, Yinlong Dai, Irmak Guzey, Siddhant Haldar, Soumith Chintala, and Lerrel Pinto. 2024. Open teach: A versatile teleoperation system for robotic manipulation. *arXiv preprint arXiv:2403.07870* (2024).

Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. 2024. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196* (2024).

Daniel Kahneman. 2011. *Thinking, fast and slow.* macmillan.

Yi Li, Yuquan Deng, Jesse Zhang, Joel Jang, Marius Memmel, Raymond Yu, Caelan Reed Garrett, Fabio Ramos, Dieter Fox, Anqi Li, et al. 2025. Hamster: Hierarchical action models for open-world robot manipulation. *arXiv preprint arXiv:2502.05485* (2025).

Zhuoling Li, Liangliang Ren, Jinrong Yang, Yong Zhao, Xiaoyang Wu, Zhenhua Xu, Xiang Bai, and Hengshuang Zhao. 2024. VIRT: Vision Instructed Transformer for Robotic Manipulation. *arXiv preprint arXiv:2410.07169* (2024).

Toru Lin, Zhao-Heng Yin, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. 2024. Twisting Lids Off with Two Hands. *CoRR* abs/2403.02338 (2024). https://doi.org/10.48550/arXiv.2403.02338

Haofei Lu, Dongqi Han, Yifei Shen, and Dongsheng Li. 2025. What Makes a Good Diffusion Planner for Decision Making?. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=7BQkXXM8Fy

Jianlan Luo, Charles Xu, Jeffrey Wu, and Sergey Levine. 2024b. Precise and Dexterous Robotic Manipulation via Human-in-the-Loop Reinforcement Learning. *arXiv preprint arXiv:2410.21845* (2024).

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris Kitani, and Weipeng Xu. 2024a. Omnigrasp: Grasping diverse objects with simulated humanoids. *Advances in Neural Information Processing Systems* 37 (2024), 2161–2184.

Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. 2023. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10895–10904.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. 2023. Orbit: A Unified Simulation Framework for Interactive Robot Learning Environments. *IEEE Robotics and Automation Letters* 8, 6 (2023), 3740–3747. https://doi.org/10.1109/LRA.2023.3270034

Soohwan Park, Hoseok Ryu, Seyoung Lee, Sunmin Lee, and Jehee Lee. 2019. Learning predict-and-simulate policies from unorganized human motion data. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–11.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. 2018. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions On Graphics (TOG)* 37, 4 (2018), 1–14.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. 2022. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)* 41, 4 (2022), 1–17.

Eduardo Pignatelli, Johan Ferret, Matthieu Geist, Thomas Mesnard, Hado van Hasselt, Olivier Pietquin, and Laura Toni. 2023. A survey of temporal credit assignment in deep reinforcement learning. *arXiv preprint arXiv:2312.01072* (2023).

Sergey Prokudin, Christoph Lassner, and Javier Romero. 2019. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4332–4341.

Martin L Puterman. 2014. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. 2021. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11488–11499.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 627–635.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. 2020. 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 16*. Springer, 581–600.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. 2024a. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–21.

Chen Tessler, Jordan Juravsky, Yunrong Guo, Yifeng Jiang, Erwin Coumans, and Xue Bin Peng. 2024b. ProtoMotions: Physics-based Character Animation. https://github.com/NVLabs/ProtoMotions/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Jungdam Won, Deepak Gopinath, and Jessica Hodgins. 2020. A scalable approach to control diverse behaviors for physically simulated characters. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 33–1.

Yan Wu, Korrawe Karunratanakul, Zhengyi Luo, and Siyu Tang. 2025. UniPhys: Unified Planner and Controller with Diffusion for Flexible Physics-Based Character Control. *arXiv preprint arXiv:2504.12540* (2025).

Sirui Xu, Hung Yu Ling, Yu-Xiong Wang, and Liang-Yan Gui. 2025. InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions. In *CVPR*.

Runyi Yu, Yinhuai Wang, Qihan Zhao, Hok Wai Tsui, Jingbo Wang, Ping Tan, and Qifeng Chen. 2025. SkillMimic-V2: Learning Robust and Generalizable Interaction Skills from Sparse and Noisy Demonstrations. *arXiv preprint arXiv:2505.02094* (2025).

Yanjie Ze, Zixuan Chen, JoÃÇo Pedro AraÃšjo, Zi-ang Cao, Xue Bin Peng, Jiajun Wu, and C Karen Liu. 2025. TWIST: Teleoperated Whole-Body Imitation System. *arXiv preprint arXiv:2505.02833* (2025).

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5745–5753.

A Humanoid model and control

Our humanoid character is based on the SMPL-X model structure [Pavlakos et al. 2019], consistent with recent work [Luo et al. 2024a; Xu et al. 2025]. It possesses $N_a = 153$ actuated degrees of freedom (DoFs), corresponding to 51 joints, each with 3 DoFs. The policy's actions $a_t \in \mathbb{R}^{N_a}$ specify target positions for Proportional-Derivative (PD) controllers at each actuated DoF.

The humanoid's body links are represented using primitive geometries (capsules and boxes). This simplified geometric representation, particularly for the hands, can generate fewer contact points compared to real human hands with soft, deformable tissue. This difference can make certain fine-grained manipulation tasks harder to reconstruct. To better approximate the contact dynamics of human soft tissue interacting with objects using rigid body simulation, we use a global friction coefficient of 1.5. This helps mitigate real-to-sim mismatch by enabling more stable grasps, such as single-handed grasping of larger objects. Furthermore, to ensure more natural and human-like finger movements, we constrain the joint limits of the hand DoFs to be within biomechanically plausible ranges, preventing uncanny configurations that can arise from the full rotational capacity of the original SMPL-X model.

In addition, we model all objects with a fixed mass of $1[KG]$ and set the restitution at 0.7.

B Retargeting

The goal in this paper is not to reconstruct the entire GRAB dataset, but rather to learn full-body-manipulation skills. As such, we aim to utilize as much data as possible, and focus on a single humanoid morphology.

Transfering motion across morphologies isn't a trivial problem. Not only does the body-size ratio differ, but also the relative lengths of various body parts with respect to the character's height. The simplest approach for transferring motion between characters with similar (same structure) morphologies is by applying the DoF rotations from the original skeleton onto the new target skeleton. While simple, and this works nicely for learning locomotion, this does not apply to the object.

Unlike the humanoid's body parts, the object is not a fixed joint within the kinematic tree. Changing the humanoid shape results in a misalignment with the object. For example, a taller human lifting an object might lift it higher.

To tackle this, we combine an explicit object retargeting strategy with the robustness of reinforcement learning within our physics simulation. The object retargeting step, illustrated in Figure 3, first corrects for the large translational discrepancies caused by the change in body shape, bringing the object approximately to the correct position relative to our character. From this more feasible starting point, the RL-trained policy, by interacting within the physics simulation, can then overcome smaller residual imperfections and learn to successfully reconstruct the nuances of the original human-object interaction.

When a reference motion (performed by a subject with a specific body shape, denoted "original") is applied to the target humanoid (which may have a different "target" shape), we adjust the object's trajectory to maintain interaction consistency. This simple approach preserves the relative positioning between the contacting hand(s) and the object. Leveraging the contact annotations, we determine the object translation offset p by optimizing the following objective to best preserve the original contact relationships:

$$p^* = \mathrm{argmin}_p \sum_{j \in \mathrm{ContactLinks}} \left| \left(\hat{c}^{\mathrm{original}}_{j,t} - \hat{c}^{\mathrm{obj}}_{j,t}\right) - \left(\hat{c}^{\mathrm{target}}_{j,t} - \left(\hat{c}^{\mathrm{obj}}_{j,t} + p\right)\right) \right|^2$$

where $\hat{c}^{\mathrm{obj}}_{j,t}$ is the location joint j is in contact with the object mesh at time t. This objective aims to find an object translation offset p that maintains the same distance from bodies-in-contact to their corresponding contact coordinate on the object mesh. The retargeted object position is then $\hat{p}^{\mathrm{obj,retargeted}}_t = \hat{p}^{\mathrm{obj}}_t + p^*$.

Finally, as our goal is to learn manipulation behaviors and not necessarily successfully reconstruct the entire dataset, we identify and remove sequences exhibiting two primary issues: **(1) Non-hand interactions.** Motions where crucial object interactions involve body parts other than the hands (e.g., the person places sunglasses on its face), which are beyond the scope of our hand-centric manipulation focus. **(2) Retargeting instability.** Instances where the retargeting process might introduce discontinuities. We filter out motions if the retargeted object's center of mass "jumps" by more than 10cm between any two consecutive frames, as this often indicates an unstable or physically implausible retargeted interaction.

After this data processing and filtering pipeline, our training set comprises 1007 motion sequences, and our test set (derived from subject 10 in GRAB) contains 141 sequences. By closely tracking these detailed human demonstrations, π_{track} learns to overcome remaining imperfections in the reference data and to master intricate and physically nuanced behaviors, such as coordinated bimanual operations, precise tool usage, and maintaining stable object control throughout extended manipulation sequences. We provide additional technical details in the supplementary material.

C Rewards

Our reward is defined as a multiplicative reward with the following components:

Error on the human global translation: $r^{ht} = e^{-100 \cdot ||\hat{p}_t - p_t||}$.

Error on the human global rotation: $r^{hr} = e^{-2 \cdot <\hat{\theta}_t, \theta_t>}$ where $< \cdot, \cdot >$ is the quaternion difference.

Error on the human velocity: $r^{hv} = e^{-0.2 \cdot ||\hat{v}_t - v_t||}$

Error on the human angular velocity: $r^{hw} = e^{-0.02 \cdot ||\hat{\omega}_t - \omega_t||}$

Error on the human energy: $r^{pow} = e^{-0.00002 \cdot |\text{dof force}_t \cdot v_t|}$

Error on the finger contact forces:
$r^{penetration} = e^{-0.00001 \cdot \Pi_{j \in \text{contact bodies}} \text{contact force}_j}$

Contact rewards:

- Pre-contact:
  - Translation error: We compute the combined error across all future reference bodies in contact.

$$\Pi_{j \in \text{in contact in the future}} e^{-100(\hat{p}^j_t - p^j_t || + |||| \hat{p}^j_t - \hat{p}^{\mathrm{obj}}_t || - || p^j_t - p^{\mathrm{obj}}_t ||)}$$

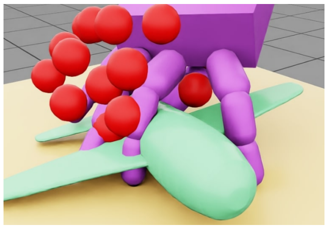


Figure 9: Penetation issues when improper reward and physics is used.

Which encourages future bodies in contact to not only maintain similar position to the reference data, but also with respect to the object. This helps prevent dynamic motion when close to contact, which may push the object far away. In addition, we use a normal cosine similarity component $\Pi_{j \in \text{in contact in the future}}(1 - \text{normal similarity}_t^j)/2$ which compares the surface normal of the closest point for each joint in the reference with that in the simulation.

- During contact we reward on the distance error for bodies that should be in contact. This encourages them to minimize the distance towards the object. A body part that is in contact is marked as distance 0. $\Pi_{j \in \text{should have contact}} e^{-2 \cdot d_t^j}$ where d is the distance to the closest point on the object surface.
- Post-contact:
  - Translation error: We compute the combined error across all previous reference bodies in contact.

$$\Pi_{j \in \text{in contact in the past}} e^{-100(\hat{p}_t^j - p_t^j|| + |||| \hat{p}_t^j - \hat{p}_t^{\text{obj}}|| - ||p_t^j - p_t^{\text{obj}}||)}$$

This encourages the agent to smoothly release the object and ensuring it remains in a similar position/rotation as the reference.

Error on the object global translation: $r^{ht} = e^{-100 \cdot ||\hat{p}_t - p_t||}$.

Error on the object global rotation: $r^{hr} = e^{-1 \cdot <\hat{\theta}_t, \theta_t>}$ where $< \cdot, \cdot >$ is the quaternion difference.

We found that the reward and humanoid design were crucial to preventing penetrations, especially with small objects. A full hand grasp applies contact forces on all sides. These forces can be very high when the humanoid approaches with full body dynamics. This tends to result in penetration issues. These penetrations are not just visually unpleasing but also pose a problem in reconstructing the motions. For example, we observed issues such as a finger getting stuck inside an object. The object then follows the hand and the agent is unable to let go of it.

Reducing finger forces and ensuring the approach is smooth and contact forces remain within reasonable ranges – these design choices help prevent penetrations such as shown in Figure 9.