Learning Robust and Generalizable Interaction Skills from Sparse and Noisy Demonstrations

RUNYI YU*, The Hong Kong University of Science and Technology, China
YINHUAI WANG*†, The Hong Kong University of Science and Technology, China
QIHAN ZHAO*, The Hong Kong University of Science and Technology, China
HOK WAI TSUI, The Hong Kong University of Science and Technology, China
JINGBO WANG, Shanghai Aritificial Intelligence Laboratory, China
PING TAN‡, The Hong Kong University of Science and Technology, China
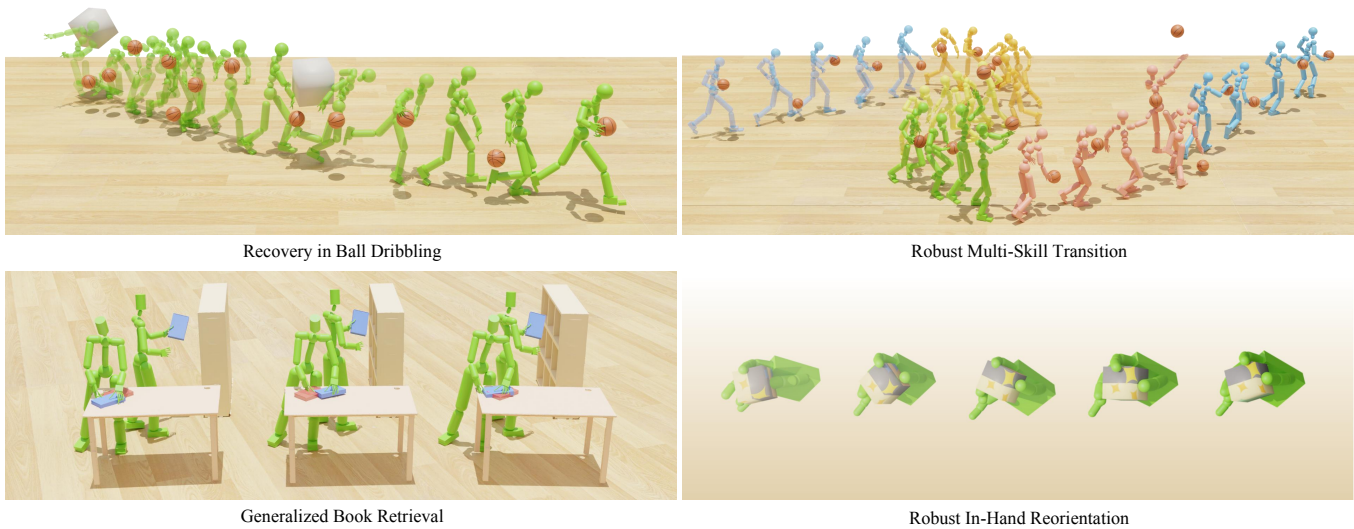QIFENG CHEN‡, The Hong Kong University of Science and Technology, China

Fig. 1. Our framework enables physically simulated robots to learn robust and generalizable interaction skills from sparse demonstrations: (top left) Learning sustained and robust dribbling from a single, brief demonstration; (top right) acquiring robust skill transitions from fragment skill demonstrations; (bottom left) generalizing book grasping to varied poses from one demonstration; and (bottom right) learning to reorientate a cube from a single grasp pose.

*Equal contribution.
†Project lead.
‡Corresponding author.

Authors' addresses: Runyi Yu, The Hong Kong University of Science and Technology, China, ingrid.yu@connect.ust.hk; Yinhuai Wang, The Hong Kong University of Science and Technology, China, yinhuai.wang@connect.ust.hk; Qihan Zhao, The Hong Kong University of Science and Technology, China, qihan.zhao@outlook.com; Hok Wai Tsui, The Hong Kong University of Science and Technology, China, hwtsui@connect.ust.hk; Jingbo Wang, Shanghai Aritificial Intelligence Laboratory, China, wangjingbo1219@gmail.com; Ping Tan, The Hong Kong University of Science and Technology, China, pingtan@ust.hk; Qifeng Chen, The Hong Kong University of Science and Technology, China, cqf@ust.hk.

We address a fundamental challenge in Reinforcement Learning from Interaction Demonstration (RLID): demonstration noise and coverage limitations. While existing data collection approaches provide valuable interaction demonstrations, they often yield sparse, disconnected, and noisy trajectories that fail to capture the full spectrum of possible skill variations and transitions. Our key insight is that despite noisy and sparse demonstrations, there exist infinite physically feasible trajectories that naturally bridge between demonstrated skills or emerge from their neighboring states, forming a continuous space of possible skill variations and transitions. Building upon this insight, we present two data augmentation techniques: a Stitched Trajectory Graph (STG) that discovers potential transitions between demonstration skills, and a State Transition Field (STF) that establishes unique connections for arbitrary states within the demonstration neighborhood. To enable effective RLID with augmented data, we develop an Adaptive Trajectory Sampling (ATS) strategy for dynamic curriculum generation and a historical encoding mechanism for memory-dependent skill learning. Our approach enables robust skill acquisition that significantly generalizes beyond the reference demonstrations. Extensive experiments across diverse interaction tasks demonstrate substantial improvements over state-of-the-art methods in terms of convergence stability, generalization capability, and recovery robustness.

1 INTRODUCTION

Robot-object interaction skills are fundamental to numerous applications, ranging from character animation to robotic manipulation [Fu et al. 2024b,a; Gao et al. 2024a; Wang et al. 2024b; Xiao et al. 2024; Xu and Wang 2024; Zhang et al. 2023b]. Recent advancements in reinforcement learning from interaction demonstration (RLID) have yielded promising results in acquiring these complex skills [Wang et al. 2023a, 2024c; Zhang et al. 2023a]. By focusing on robot-object state transitions, a unified learning framework has been established, enabling the acquisition of versatile interaction skills from diverse human demonstrations efficiently. However, while current demonstration collection methods provide rich interaction examples, the captured trajectories are usually noisy and sparse - only capture a limited subset of possible skill variations rather than the full spectrum of interaction patterns [Fan et al. 2024, 2023; Jiang et al. 2023; Kim et al. 2024; Liu et al. 2022; Menolotto et al. 2020; Taheri et al. 2020; Wang et al. 2024c; Zhang et al. 2024b]. Therefore, developing methods to acquire robust and generalizable interaction skills from sparse and noisy demonstrations is of particular importance.

In this work, we present a novel data augmentation and training system built upon RLID that significantly enhances its capabilities in handling imperfect demonstrations, achieving superior convergence stability, robustness to perturbations, and generalization performance. Our key insight is that despite noisy and sparse demonstrations, there exist infinite physically feasible trajectories that naturally bridge between demonstrated skills or emerge from their neighboring states, forming a continuous space of possible skill variations and transitions. Building upon this insight, we develop a comprehensive data augmentation framework to fully identify these uncaptured skill patterns. The framework consists of two core components: a Stitched Trajectory Graph (STG) that discovers potential transitions between demonstration skills, and a State Transition Field (STF) that establishes unique connections for arbitrary states within the demonstration neighborhood. To facilitate effective STF learning through RLID, we introduce an Adaptive Trajectory Sampling (ATS) strategy to ensure balanced learning of hard samples, complemented by a pre-trained history encoder for memory-dependent skill learning.

Given sparse and noisy demonstrations, our method not only acquires intended interaction skills but also achieves robust recovery capabilities from error states within the demonstration neighborhood. Furthermore, our approach masters unseen bridging transitions between demonstrated skills, enabling robust and smooth skill switching. This demonstrates the potential of our method in

enriching the coverage of interaction and manipulation patterns that are typically challenging to capture during data collection.

Extensive experiments across diverse datasets, including BallPlay-M [Wang et al. 2024c] and ParaHome [Kim et al. 2024], demonstrate substantial improvements over state-of-the-art approaches. Our method achieves near-perfect success rates with 40-50% improvement and enhances generalization performance by over 35% compared to existing methods. Comprehensive ablation studies and case analyses further validate the effectiveness of each proposed component. We encourage readers to visit our project website for video demonstrations: **https://anonymous202501.github.io**

2 RELATED WORK

2.1 Imitation Learning in Character Animation

In recent years, the field of learning physics-based character skills from demonstrations has witnessed remarkable advancements [Bae et al. 2023; Braun et al. 2023; Dou et al. 2023; Hassan et al. 2023; Liu and Hodgins 2018; Luo et al. 2023; Pan et al. 2024; Park et al. 2019; Peng et al. 2018, 2022; Sferrazza et al. 2024; Tessler et al. 2023; Wang et al. 2023a, 2024c; Zhang et al. 2023b,a]. Broadly, these methods can be categorized into two types: locomotion and interaction.

Locomotion. Recently, reinforcement learning [Kaelbling et al. 1996] within physics-based simulation environments [Makoviychuk et al. 2021], guided by imitation reward functions, emerged as the mainstream approach for humanoid skill acquisition. This shift has been instrumental in both character animation [Peng et al. 2018, 2022, 2021] and the development of robust gaits for real-world robots [Fu et al. 2024b; He et al. 2024a,b,c; Zhang et al. 2024c]. [Peng et al. 2018, 2022, 2021] introduced classic aligned imitation reward functions and unaligned adversarial imitation reward functions [Ho and Ermon 2016] to learn locomotion skills. Further research has applied imitation rewards to motion-tracking [Luo et al. 2023] and conditional control [Dou et al. 2023; Tessler et al. 2023]. These seminal works inspired subsequent research that leverages locomotion priors for learning diverse interaction tasks [Dou et al. 2023; Hassan et al. 2023; Liu et al. 2024; Peng et al. 2022; Tessler et al. 2024, 2023; Xiao et al. 2024], such as playing tennis [Zhang et al. 2023b], climbing ropes [Bae et al. 2023], and grasping [Luo et al. 2024].

Interaction. A significant body of research in Human-Object Interaction (HOI) has emphasized non-physical generative approaches [Jiang et al. 2023, 2024b; Li et al. 2023a, 2025, 2023b; Starke et al. 2020, 2021; Wang et al. 2024a; Xu et al. 2023a, 2024; Yang et al. 2025]. Despite their advantages in multimodal integration and scalability, these methods inherently lack physical authenticity and necessitate extensive training data. Recent works have attempted to extend the success of imitation learning in locomotion to interactive skill acquisition, forming an emerging paradigm we term as Reinforcement Learning from Interaction Demonstration (RLID). Zhang et al. [Zhang et al. 2023a] introduced interaction graph for learning multi-character interactions and retargeting. Chen et al. [Chen et al. 2024] developed a hierarchical policy learning framework that leverages human hand motion data to train object-centric dexterous robot manipulation. Most relevant to our work, SkillMimic [Wang et al. 2023a, 2024c] proposed a scalable framework for RLID with a unified
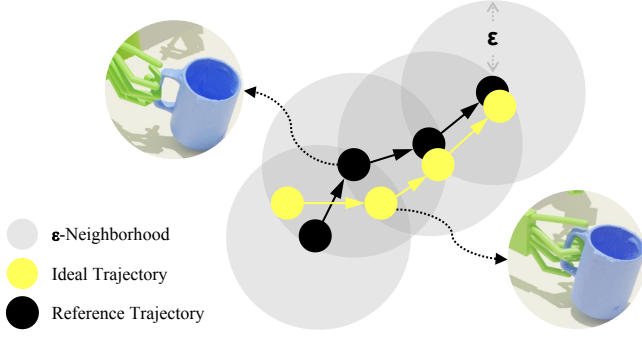
Fig. 2. Given a degraded reference trajectory containing physically unreachable state transitions, perfect trajectory reconstruction becomes impossible. The goal is to learn a set of ideal trajectories that are both physically feasible and satisfy reconstruction thresholds. These ideal trajectories must exist within an ε-neighborhood of the reference trajectory.

interaction imitation reward, enabling the acquisition of complex basketball skills such as dribbling and shooting, while demonstrating the reusability of these learned interaction skills.

2.2 Data Augmentation for Motion Data

Data augmentation for motion capture has been a long-standing challenge in character animation. Early approaches relied on motion graphs [Kovar et al. 2002; Lee et al. 2002; Zhao and Safonova 2009] to synthesize continuous animations by concatenating motion fragments. These methods typically identify similar motion frames through pose matching and resolve discontinuities via motion blending techniques [Kovar and Gleicher 2003, 2004]. While such approaches have achieved remarkable success in locomotion synthesis, their extension to human-object interaction (HOI) scenarios remains challenging. Motion graphs inherently require a comprehensive motion database to enable effective transitions between motion segments. However, in HOI contexts, the introduction of manipulated objects dramatically expands the interaction space, making it prohibitively expensive to capture sufficient data covering all possible transition scenarios.

Recent years have witnessed the emergence of rule-based data augmentation methods for robot-object trajectories [Gao et al. 2024b; Garrett et al. 2024; Jiang et al. 2024a; Mandlekar et al. 2023; Pumacay et al. 2024; Zhang et al. 2024a]. While these approaches have shown promise in expanding manipulation datasets, they face fundamental limitations when handling noisy demonstrations or bridging sparse motion segments in the manipulation space. In contrast, RLID [Wang et al. 2023a, 2024c] has demonstrated remarkable tolerance to data noise, and generative adversarial imitation learning (GAIL) [Ho and Ermon 2016] with random state initialization [Andrychowicz et al. 2020; Hwangbo et al. 2019] has proven effective in learning generalized transitions between sparse motion segments in locomotion tasks [Peng et al. 2022]. However, the successful application of GAIL to interaction imitation remains an open challenge, since interactions require more fine-grained guidance compared to locomotion, whereas GAN rewards tend to be coarse-grained [Wang et al. 2024c].

3 PRELIMINARIES ON RLID

Reinforcement Learning from Interaction Demonstration (RLID) views the learning of the manipulation task as learning underlying robot-object state transitions [Wang et al. 2024c], which is typically defined by a reference trajectory $\mathcal{A} : \{\hat{s}_0, ..., \hat{s}_T\}$ where T represents the trajectory length, $\hat{s}_t$ represents the kinematics of both the robot and objects. The state transitions evolve through the interplay of a learned policy $\pi(a_t|s_t)$ and a deterministic physics simulator $f(s_{t+1}|a_t, s_t)$. The policy is parameterized as a Gaussian distribution to enable stochastic exploration, where the mean is generated by a neural network $\phi(s_t)$ that maps states to actions, while maintaining a fixed variances Σ. Formally, we have $a_t \sim \mathcal{N}(\phi(s_t), \Sigma)$. We can also rewrite s_{t+1} as a stochastic variable:

$$s_{t+1} \sim P(\cdot|\phi, s_t, f). \tag{1}$$

To learn the target state transitions, a unified interaction imitation reward [Wang et al. 2024c] is used to measure the similarity between the generated robot-object state and the reference:

$$r_t = S(s_{t+1}, \hat{s}_{t+1}) = r_t^b * r_t^o * r_t^{rel} * r_t^{cg}, \tag{2}$$

which integrates four normalized sub-rewards: body states (r_t^b), object states (r_t^o), robot-object relative positions (r_t^{rel}), and contacts (r_t^{cg}). The integrated reward r_t is bounded in [0,1], enabling consistent scaling across diverse demonstrations. During RLID training, the robot and object states are initialized from the reference $\hat{s}_i$ [Peng et al. 2018], where i is randomly sampled from $[0, T-1]$.

To handle diverse transition patterns, we adopt the conditioning mechanism from [Wang et al. 2024c] by introducing a condition variable c into the policy formulation $\pi(a_t|s_t, c)$. This variable can encode various levels of information, from high-level skill labels in basketball tasks to fine-grained target states for tracking models, enabling a single policy to learn from heterogeneous demonstrations.

4 METHOD

4.1 Problem Definition

Given a noisy reference trajectory $\mathcal{A} : \{\hat{s}_0, ..., \hat{s}_T\}$ containing both degraded and missing states, where each state $\hat{s}_t = [\hat{s}_t^r, \hat{s}_t^o]$ comprises both robot state $\hat{s}_t^r$ and object state $\hat{s}_t^o$. States are masked with $\mathcal{M}$ when missing from the trajectory. Our goal is to learn robust interaction skills that can generate physically feasible trajectories while maintaining similarity to the available reference states. Formally, we aim to learn a set $\mathcal{S}$ of feasible trajectories where each $\mathcal{A}^* \in \mathcal{S}$ maximizes the expected return $\mathcal{R}(\pi)$ while satisfies the following constraints:

$$\mathcal{A}^* = \{s_i^*, s_{i+1}^*, ..., s_T^*\}, \ i \in 0, ..., T, \tag{3}$$

$$\forall t \in i, ..., T-1 : (s_t^*, s_{t+1}^*) \in \mathcal{F}, \tag{4}$$

$$\forall t \in i, ..., T : \mathcal{M}_t(|s_t^* - \hat{s}_t|) \le \varepsilon. \tag{5}$$

Here, $\mathcal{F}$ denotes the set of physically plausible state transitions, encompassing both robot-object interaction dynamics and physical constraints. $\mathcal{M}_t \in \{0, 1\}$ is a binary mask indicating whether the reference state at time t is available ($\mathcal{M}_t = 1$) or missing ($\mathcal{M}_t = 0$), and ε defines the tolerance bounds for each dimension of the robot and object states. Each trajectory $\mathcal{A}^*$ can start from any time step
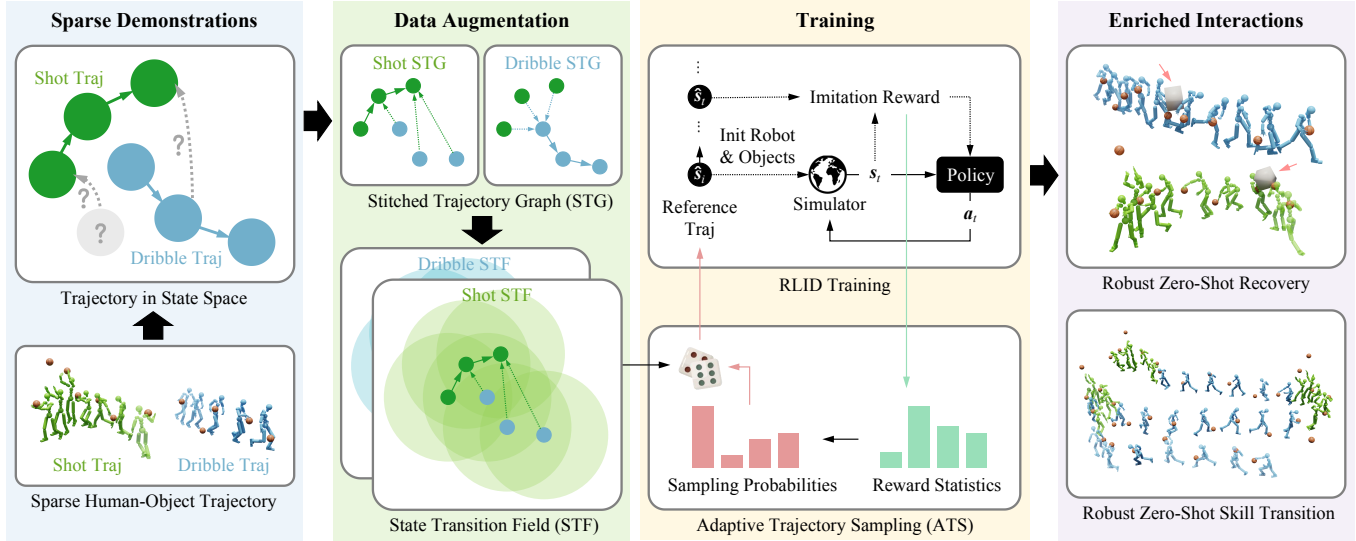
Fig. 3. Given sparse demonstrations (e.g., two short trajectories of Shot and Dribble), there exist infinite valid but uncaptured trajectories that can either bridge between them or emerge from their neighboring states (illustrated by question marks). Our method discovers and learns these potential trajectories through three key steps: (1) constructing a Stitched Trajectory Graph (STG) to identify possible transitions between demonstrations, (2) expanding STG into a State Transition Field (STF) that establishes connections for arbitrary states within the demonstration neighborhood, and (3) learning a skill policy via Adaptive Trajectory Sampling (ATS) and Reinforcement Learning from Interaction Demonstrations (RLID). This enables robust skill transition and generalization far beyond the original sparse demonstrations which contain no instances of skill transitions or error recovery patterns.

i and any state within an ε-neighborhood of $\hat{s}_i$, progressively converging towards the reference trajectory until time T. The set $\mathcal{S}$ thus represents all physically feasible trajectories in this neighborhood, characterizing the robustness and generalization of an ideal skill.

4.2 Motivation and Method Overview

While the basic RLID method [Wang et al. 2024c] relies on state initialization from the reference trajectory [Peng et al. 2018], it becomes problematic when handling data degradation. Unlike locomotion imitation, interaction tasks are highly sensitive to data perturbations - even a 2cm deviation between finger and objects may cause catastrophic failures. Fig. 2 illustrates a typical error pattern in reference data. When data degradation exists around time step i, the learning of the entire state transition chain may break around time step i, resulting in near-zero success rates despite the policy converging well on other demonstration segments.

From Eq. 5, we know that the target trajectory set $\mathcal{S}$ lies within the ε-neighborhood of the degraded reference trajectory $\mathcal{A}$, as illustrated in Fig. 2. Therefore, random initialization within the entire ε-neighborhood theoretically ensures complete coverage of states in $\mathcal{S}$, potentially providing better escape from local optima compared to initialization from fixed erroneous states. This insight motivates our data augmentation framework that establishes directed transitions for every state in the neighborhood, naturally forming a field of state transitions - hence termed State Transition Field (STF). Moreover, as implied by Eq. 1, increasing sampling frequency around challenging segments enhances the probability of discovering valid trajectories in $\mathcal{S}$. By allocating larger sampling weights to harder state transitions, we can better address the "chain break" problem

and improve the success rate of complete trajectory execution. We term this method as Adaptive Trajectory Sampling (ATS).

Given the ability to handle noisy and incomplete data, we further augment the training data by stitching different trajectories to form a graph structure termed Stitched Trajectory Graph (STG). STF is then built upon STG to provide broader state coverage. For demonstrations with different condition labels, we construct separate STFs for each condition c, where trajectories sampled from STF inherit the corresponding condition. During training, ATS samples reference trajectories from these STFs, allowing the policy to learn diverse state transition patterns conditioned on c. Fig. 3 illustrates this process using basketball skill learning as an example. The following subsections detail these technical components.

4.3 State Transition Field

A straightforward way for neighborhood sampling is to add noise ε to the basic RLID initialization when starting from s_t [Liu et al. 2010; Peng et al. 2022]. However, this leads to convergence issues both theoretically and empirically. Specifically, neighborhoods of different states may have significant overlap, especially when states are close or the neighborhood range is large. In such cases, a state s_{new} may simultaneously belong to multiple reference state neighborhoods, leading to convergence challenges due to the non-unique mapping of state transitions. Therefore, unique transition directions must be established for each neighborhood state to ensure convergence.

Moreover, when neighborhoods are large, transitions from the border to the center may be physically infeasible in a single simulation step. We resolve this by inserting masked states between distant points as potential missing data to be inpainted. These masked states

contain no predefined values, serve purely as temporal buffers for bridging distant transitions and are excluded from reward computation. This essentially constructs missing data patterns that can be repaired through RLID.

The unique transition directions for all states in the neighborhood form a field of state transitions, which we term State Transition Field (STF). During training, we randomly sample trajectories from STF for RLID training. In the following, we detail the process of sampling a trajectory from STF.

4.3.1 ε-Neighborhood State Initialization (ε-NSI). Given a reference trajectory $\mathcal{A} = \{\hat{s}_0, ..., \hat{s}_T\}$, we randomly select a time i and sample a new state s_{new} uniformly from the ε-neighborhood of the reference state $\hat{s}_i$ as the initial state of the sampled trajectory.

4.3.2 Connection Rules. We then compute the similarity metric between s_{new} and all reference states in $\mathcal{A}$, identifying the state $\hat{s}_j$ that exhibits maximum similarity:

$$\hat{s}_j = \arg\max_{s \in \mathcal{A}} S(s_{\text{new}}, s). \tag{6}$$

Based on the similarity score between s_{new} and $\hat{s}_j$, we determine the required number N of masked states $s_\varnothing$ to ensure feasible state transitions. The sampled trajectory is constructed as:

$$\{s_{\text{new}}, \underbrace{s_\varnothing, ..., s_\varnothing}_{N}, \hat{s}_j, ..., \hat{s}_T\}, \tag{7}$$

where s_{new} serves as the initialization state. The detailed computation of similarity metrics and masked node numbers is provided in the supplementary material.

4.4 Stitched Trajectory Graph

For sparse demonstrations, there often exist potential connections between them that were simply not captured during data collection. We can artificially construct these connections between demonstrations and use masks to indicate the missing data. Benefiting from STF's capability in handling noisy and incomplete data, these artificially introduced "noise" and "missing data" through manual stitching can be effectively repaired. This stitching approach effectively expands the coverage of the demonstration space while maintaining the inherent structure of the original demonstrations.

Consider a trajectory $\mathcal{A}$ representing skill A, and a set $\mathcal{B}$ containing all states from trajectories of other skills. We posit that all states in $\mathcal{B}$ potentially have valid transitions to skill A, even though these transitions were not captured during data collection. By stitching these potential trajectories with $\mathcal{A}$, we are essentially construct a Stitched Trajectory Graph (STG) of skill A, denoted as $\mathcal{A}^\dagger$. Specifically, for each state in $\mathcal{B}$, we employ similar connection rules as described in Sec. B.4 to construct its path to trajectory $\mathcal{A}$. Notably, we exclude connections for states in $\mathcal{B}$ that are too distant from $\mathcal{A}$. Once the STG $\mathcal{A}^\dagger$ is constructed, we use it to replace the original reference trajectory $\mathcal{A}$ for subsequent STF data augmentation, trajectory sampling, and RLID training. Fig. 3 shows an example.

4.5 Adaptive Trajectory Sampling

To improve performance on hard samples and address the "chain break" problem, we use Adaptive Trajectory Sampling (ATS) to adaptively adjust sampling weights based on sample difficulty, which is similar to [Won and Lee 2019]. When initialized from state $\hat{s}_i$, the clip $\mathcal{A}_i = \{\hat{s}_i, ..., \hat{s}_T\}$ will be used for training. We define the sampling probability for clip $\mathcal{A}_i$ as p_i, formulated by:

$$p_i = \frac{e^{-\lambda_s * \bar{r}_i}}{\sum_{j=0}^{T-1} e^{-\lambda_s * \bar{r}_j}}, \quad \bar{r}_i = \frac{1}{T-i} \sum_{t=i}^{T-1} r_t, \tag{8}$$

where r_t is defined in Eq. 2, $\bar{r}_i$ is the average reward per frame, which quantifies the reconstruction quality when initializing from state $\hat{s}_i$. $\lambda_s \in [0, \infty)$ is a coefficient that controls the trade-off between uniform sampling ($\lambda_s = 0$) and difficulty-based sampling ($\lambda_s > 0$).

We now describe the complete STF sampling process integrated with ATS. Given an STF built upon STG:

- First, with probability p_e, we decide whether to sample the centroid state from external state set $\mathcal{B}$ or from the original trajectory $\mathcal{A}$. For the former case, we uniformly sample from $\mathcal{B}$. For the latter case, we select the centroid state according to the probabilities computed by ATS.
- Once a centroid state is selected, with probability p_n, we either sample a starting state from its neighborhood (NSI), or use it directly as the starting state.

During multi-skill learning, skills also vary in difficulty. ATS can be similarly applied to ensure balanced learning across skill classes.

4.6 History Encoder

Policies lacking temporal or historical context cannot execute memory dependent behaviors, such as determining the ball-holding duration before passing. These state transitions cannot be determined solely by the current state, as similar states in the reference trajectory may lead to different transitions at different times. This ambiguity can prevent basic RLID from converging. While phase or temporal encoding [Peng et al. 2018] can address this issue, they require manual specification, which becomes particularly challenging when dealing with multi-skill transitions. We propose History Encoder (HE) that captures temporal dependencies in a data-driven manner, operating autonomously without manual phase specifications, enabling flexible skill transitions at arbitrary states.

Formally, given a sequence of k previous states $s_{t-k}, ..., s_{t-1}$, the HE θ generates a compact historical embedding:

$$h_t = \theta(s_{t-k}, ..., s_{t-1}). \tag{9}$$

The policy network then takes this historical context into account:

$$a_t \sim \pi(\cdot | c, s_t, h_t). \tag{10}$$

To ensure stable training, we pre-train HE using behavioral cloning and freeze its parameters during RLID training. This compact history embedding prevents overfitting and alleviates PPO convergence issues that might arise from high-dimensional history observations. Details are provided in the appendix.

5 EXPERIMENT

5.1 Experimental Setup

Our study employs Isaac Gym [Makoviychuk et al. 2021] as the physics simulation platform. All training procedures are executed

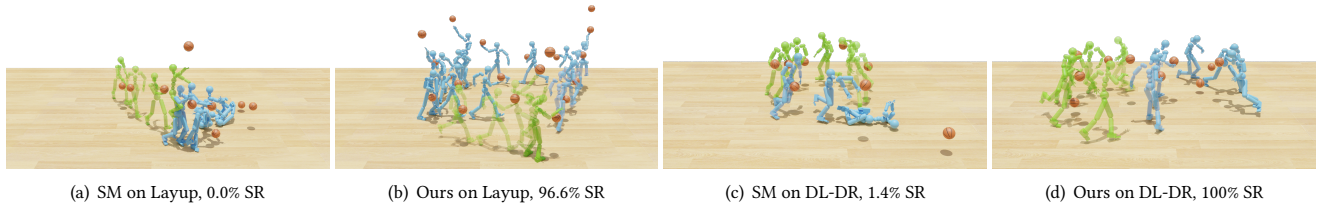| (a) SM on Layup, 0.0% SR | (b) Ours on Layup, 96.6% SR | (c) SM on DL-DR, 1.4% SR | (d) Ours on DL-DR, 100% SR |

Fig. 4. Qualitative comparison on BallPlay-M. Blue trajectories in (a,b) indicate executions beyond the reference Layup data length. In (c,d), green and blue trajectories represent dribbling left (DL) and dribbling right (DR) respectively, demonstrating skill transition not present in the reference data.

| (a) SM on Pour-Kettle-Cup | (b) SM on Stand-Chair |

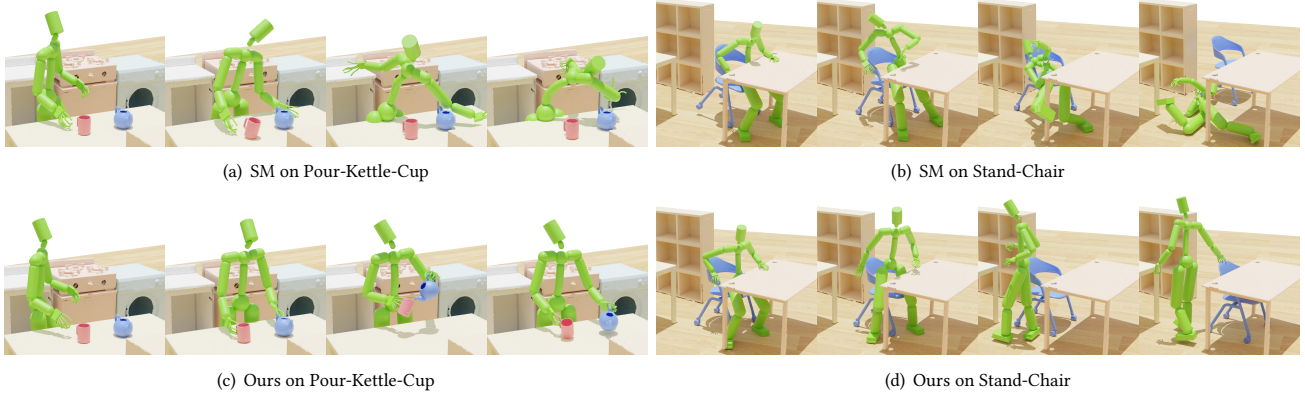| (c) Ours on Pour-Kettle-Cup | (d) Ours on Stand-Chair |

Fig. 5. Qualitative comparison on ParaHome. Humanoid performing (a,c) tea-pouring and teapot placement, (b,d) standing and chair-pushing sequences.

on a single NVIDIA RTX 4090 GPU, leveraging 2048 parallel environments. Across all experiments, the PD controller and simulation operate at a frequency of 120 Hz, while the policy is sampled at 60 Hz. We use the Proximal Policy Optimization algorithm (PPO) [Schulman et al. 2017] to optimize the policy. The simulated humanoid model replicates the kinematic tree and degrees of freedom (DoF) configurations in the demonstration dataset. Detailed hyperparameter settings are available in the appendix.

For evaluation, we consider the following four metrics:

Success Rate (SR): the percentage of successful skill executions when initialized from the reference state of the current skill. For BallPlay-M [Wang et al. 2024c], we consider a skill execution successful if it can be performed continuously for 10 seconds. For Para-Home [Kim et al. 2024], Success is defined as accurately reproducing the demonstrated interaction.

Skill Transition Success Rate (TSR): the percentage of successful target skill executions when initialized from reference states of other skills.

ε-Neighborhood Success Rate (εNSR): this metric evaluates robustness and generalization capabilities by measuring the success rate when initializing from states sampled within an ε-neighborhood of the reference trajectory.

Normalized Reward (NR): we compute the average reward per frame using NR = $\frac{1}{T} \sum_{t=0}^{T-1} \bar{r}_t$, where $\bar{r}_t$ is defined in Eq. **??** and T represents the length of the reference trajectory.

All metrics are averaged over 10,000 random trials to ensure statistical reliability.

5.2 Evaluation on BallPlay-M

Dataset and Setup. BallPlay-M [Wang et al. 2024c] is a human-basketball interaction dataset containing diverse basketball skills. We select 5 representative skills: Dribble-Forward (DF), Dribble-Left (DL), Dribble-Right (DR), Layup, and Shot, each represented by a 1-3 second clip. The skill labels are one-hot encoded as conditions c. All skills are trained using a unified policy network on a single NVIDIA RTX 4090 GPU for over 1.3 billion samples (~24 hours).

Methods. We compare our method against two representative baselines: (1) SkillMimic (SM) [Wang et al. 2024c], a state-of-the-art RLID method, and (2) DeepMimic (DM) [Peng et al. 2018], a classic locomotion imitation learning approach adapted for RLID following SM's implementation [Wang et al. 2024c]. For fair comparison, both baselines are trained with identical setup as our method. We further augment these baselines with ε-Neighborhood State Initialization (ε-NSI), denoted as SM+ε-NSI and DM+ε-NSI respectively. Finally, we implement our full method on both baselines, denoted as SM+Ours and DM+Ours. The dimension of history embedding is 3.

Quantitative Analysis. As shown in Tab. 1, baseline methods achieve satisfactory performance on dribble skills (DF, DL, DR) but struggle with scoring skills (Layup, Shot). This performance gap stems from the incomplete state transition loops in reference data for scoring skills (visualized in Fig. 1). Besides, baseline methods show limited skill transition capability due to the lack of skill transition demonstrations in the reference data.

Table 1. Quantitative comparison on BallPlay-M. The neighborhood range ε for εNSR test is consistent with training settings

| Method | SR↑ (%) / εNSR↑ (%) / NR↑ | | | | | | TSR↑ (%) | | | | |
	DF	DL	DR	Layup	Shot	Avg.	DL-DR	DF-DR	DF-Shot	Layup-DF	Avg.
DM	89.2 / 38.5 / 0.09	70.4 / 24.5 / 0.10	87.5 / 26.8 / 0.06	0.0 / 0.0 / 0.18	0.0 / 0.0 / 0.12	49.4 / 18.0 / 0.11	1.6	17.1	0.0	50.2	17.2
DM + ε-NSI	96.2 / 56.3 / 0.10	76.5 / 38.7 / 0.11	81.8 / 29.5 / 0.06	1.0 / 0.5 / 0.20	0.0 / 0.0 / 0.11	51.1 / 25.0 / 0.12	2.9	14.0	0.0	46.1	15.8
DM + Ours	88.6 / 53.8 / 0.08	89.4 / 55.9 / 0.08	92.6 / **54.2** / 0.10	**99.1** / 50.9 / 0.12	0.0 / 0.0 / 0.06	73.9 / 43.0 / 0.088	93.6	87.0	0.0	72.1	63.2
SM	97.3 / 40.3 / **0.64**	71.9 / 29.4 / 0.49	97.8 / 22.9 / 0.35	0.0 / 0.0 / **0.64**	2.9 / 1.3 / **0.48**	54.0 / 18.8 / **0.52**	1.4	28.0	0.8	31.2	15.4
SM + ε-NSI	98.1 / 61.2 / 0.38	98.7 / **59.8** / **0.51**	97.1 / 44.9 / **0.36**	23.1 / 12.1 / 0.62	0.0 / 0.0 / 0.39	63.4 / 35.6 / 0.45	37.2	77.3	0.0	49.3	41.0
SM + Ours	**99.8 / 62.7** / 0.37	**100.0** / 42.2 / 0.42	**99.7** / 42.0 / 0.34	96.6 / **58.5** / 0.57	**100.0 / 56.0** / 0.46	**99.2 / 52.3** / 0.43	**100.0**	**99.7**	**99.8**	**99.8**	**99.8**

Table 2. Quantitative comparison on ParaHome. The neighborhood range ε for εNSR test is object-centric.

| Method | SR↑ (%) / εNSR↑ (%) / NR↑ | | | | | | | |
	Place-Pan	Place-Kettle	Place-Book	Drink-Cup	Pour-Kettle	Stand-Chair	Pour–Kettle-Cup	Avg.
SM	38.4 / 1.0 / 0.92	0.0 / 0.0 / 0.51	0.0 / 0.0 / 0.53	0.0 / 0.0 / 0.39	0.0 / 0.0 / 0.84	0.0 / 0.0 / 0.55	0.0 / 0.0 / 0.79	5.5 / 0.1 / 0.65
SM + ε-NSI	100 / 16.2 / 0.88	0.0 / 0.0 / 0.53	0.0 / 0.0 / 0.32	0.0 / 0.0 / 0.42	0.0 / 0.0 / 0.67	0.0 / 0.0 / 0.55	0.0 / 0.0 / 0.74	14.3 / 2.3 / 0.59
SM + T	51.6 / 12.6 / **0.93**	0.0 / 0.0 / 0.30	100 / 12.1 / 0.85	99.9 / 20.3 / 0.84	**100** / 21.5 / **0.95**	0.0 / 0.0 / 0.63	48.1 / 8.0 / 0.86	57.1 / 15.6 / 0.77
SM + Ours	**100 / 22.2** / 0.86	**100 / 49.9 / 0.52**	**100 / 82.4 / 0.86**	**100 / 33.9 / 0.89**	99.9 / **22.5** / 0.93	**100 / 46.6 / 0.74**	**100 / 23.1 / 0.87**	**100 / 40.1 / 0.81**

While naive ε-NSI provides moderate improvements, our full method demonstrates substantial performance gains across all metrics: +45% in average SR; +33% in average εNSR; +84% in average TSR. Although SM achieves the highest NR, demonstrating strong fitting capacity on the reference dataset, it shows unbalanced success rate and poor generalization performance. In contrast, our method not only fits the reference data well but also exhibits strong generalization and robustness to out-of-domain cases.

Qualitative Analysis. Fig. 1(a) and Fig. 4(b) demonstrate the superior robustness and generalization capabilities of our method, despite learning each skill from only a single noisy reference trajectory. Learning from just five sparse demonstrations, Fig. 1(b) showcases smooth skill transitions that never shown in the reference dataset. These visualizations, combined with the quantitative results, validate our method's capability in learning robust and generalizable interaction skills from sparse and noisy demonstrations.

5.3 Evaluation on ParaHome

Datase and Setup. ParaHome dataset [Kim et al. 2024] features diverse human-object interactions in household scenarios. We evaluate on 7 representative interaction clips: Place-Pan, Place-Kettle, Place-Book, Drink-Cup, Pour-Kettle, Stand-Chair, and Pour-Kettle-Cup, each spanning 2-5 seconds. For each clip, we train an independent policy for around 1.0 billion samples on a single GPU. To evaluate object-centric generalization, during testing, we set ε as random perturbations of object pose: -45° to 45° rotation around z-axis and up to 10cm positional offset in the xy-plane.

Methods. We maintain similar method settings as in BallPlay-M experiment. However, our full approach here excludes the STG component, as trajectories involving different objects (e.g., kettle vs. chair) cannot form meaningful connections for cross-skill learning. We additionally condition SM with reference time t as c to examine the effect of historical information, denoted as SM+T.

Table 3. Ablation study on key components. Conducted on BallPlay-M.

Method	SR↑ (%)	TSR↑ (%)	εNSR↑ (%)	NR↑
SM	53.98	18.25	15.36	**0.52**
SM + ATS	56.39	22.43	19.74	0.42
SM + HE	60.07	34.52	7.09	0.48
SM + STF	68.67	35.07	36.96	0.43
SM + STG	74.74	71.67	28.91	0.44
SM + STG + STF	77.12	73.18	45.08	0.40
SM + STG + STF + ATS	76.44	70.23	45.67	0.39
SM + STG + STF + ATS + HE (Full)	**99.21**	**99.84**	**52.27**	0.43

Quantitative Analysis. The baseline method SM largely fails on these tasks due to small-scale motions and concentrated states, making action decisions challenging without historical context. Moreover, the impact of data noise is amplified in finger-level manipulation, hindering convergence. While incorporating ε-NSI and T separately addresses some issues, our full method further enhances overall success rate, robustness, and convergence across all tasks.

Qualitative Analysis. As shown in Fig. 5, SM struggles to reconstruct these interactions under the compound challenges of demonstration inaccuracy and ambiguous state transitions. In contrast, our method achieves natural interaction imitation. Fig. 1(c) shows the generalization capabilities learned from a single demonstration.

5.4 Ablation Study

To assess each component's contribution, we perform comprehensive ablation studies on BallPlay-M following the experimental setup detailed in Sec. 5.2. Results in Tab. 3 reveals that each proposed component yields substantial performance gains. The synergistic integration of these components in our full method achieves the optimal performance, validating our design choices.

Additional experiments on data efficiency, in-hand reorientation, and locomotion skills are available in the appendix.

6 CONCLUSION

In this paper, we introduce a novel data augmentation and learning framework that fundamentally advances the learning of robust and generalizable interaction skills from sparse and noisy demonstrations. Through extensive experiments on basketball manipulations and diverse household tasks, coupled with comprehensive ablation studies, our approach demonstrates substantial improvements over state-of-the-art methods.

While our framework shows limitations with heavily corrupted demonstrations, incorporating large-scale interaction priors (e.g., training tracking policies conditioned on target robot-object states) could address these challenges. Given our framework's unique capability to extract rich manipulation patterns from sparse noisy data, it shows promise as a fundamental building block for broader applications in both animation synthesis and robotic skill acquisition.

REFERENCES

OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. 2020. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research* 39, 1 (2020), 3–20.

Jinseok Bae, Jungdam Won, Donggeun Lim, Cheol-Hui Min, and Young Min Kim. 2023. Pmp: Learning to physically interact with environments using part-wise motion priors. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–10.

Jona Braun, Sammy Christen, Muhammed Kocabas, Emre Aksan, and Otmar Hilliges. 2023. Physically plausible full-body hand-object interaction synthesis. *arXiv preprint arXiv:2309.07907* (2023).

Yuanpei Chen, Chen Wang, Yaodong Yang, and C Karen Liu. 2024. Object-centric dexterous manipulation from human motion data. *arXiv preprint arXiv:2411.04005* (2024).

Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. 2023. C-ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. 2024. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 494–504.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. 2023. ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12943–12954.

Zipeng Fu, Qingqing Zhao, Qi Wu, Gordon Wetzstein, and Chelsea Finn. 2024b. HumanPlus: Humanoid Shadowing and Imitation from Humans. *arXiv preprint arXiv:2406.10454* (2024).

Zipeng Fu, Tony Z Zhao, and Chelsea Finn. 2024a. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117* (2024).

Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. 2024a. CooHOI: Learning Cooperative Human-Object Interaction with Manipulated Object Dynamics. *arXiv preprint arXiv:2406.14558* (2024).

Jensen Gao, Annie Xie, Ted Xiao, Chelsea Finn, and Dorsa Sadigh. 2024b. Efficient Data Collection for Robotic Manipulation via Compositional Generalization. *arXiv preprint arXiv:2403.05110* (2024).

Caelan Garrett, Ajay Mandlekar, Bowen Wen, and Dieter Fox. 2024. SkillMimicGen: Automated Demonstration Generation for Efficient Skill Learning and Deployment. *arXiv preprint arXiv:2410.18907* (2024).

Mohamed Hassan, Yunrong Guo, Tingwu Wang, Michael Black, Sanja Fidler, and Xue Bin Peng. 2023. Synthesizing physical character-scene interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–9.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. 2024a. OmniH2O: Universal and Dexterous Human-to-Humanoid Whole-Body Teleoperation and Learning. *arXiv preprint arXiv:2406.08858* (2024).

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. 2024b. Learning Human-to-Humanoid Real-Time Whole-Body Teleoperation. *arXiv preprint arXiv:2403.04436* (2024).

Tairan He, Wenli Xiao, Toru Lin, Zhengyi Luo, Zhenjia Xu, Zhenyu Jiang, Jan Kautz, Changliu Liu, Guanya Shi, Xiaolong Wang, et al. 2024c. HOVER: Versatile Neural Whole-Body Controller for Humanoid Robots. *arXiv preprint arXiv:2410.21229* (2024).

Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems* 29 (2016).

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. 2019. Learning agile and dynamic motor skills for legged robots. *Science Robotics* 4, 26 (2019), eaau5872.

Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. 2021. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF international conference on computer vision*. 11107–11116.

Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He Wang, Yixin Zhu, and Siyuan Huang. 2023. Full-body articulated human-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9365–9376.

Nan Jiang, Zhiyuan Zhang, Hongjie Li, Xiaoxuan Ma, Zan Wang, Yixin Chen, Tengyu Liu, Yixin Zhu, and Siyuan Huang. 2024b. Scaling up dynamic human-scene interaction modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1737–1747.

Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. 2024a. DexMimicGen: Automated Data Generation for Bimanual Dexterous Manipulation via Imitation Learning. *arXiv preprint arXiv:2410.24185* (2024).

Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. 1996. Reinforcement learning: A survey. *Journal of artificial intelligence research* 4 (1996), 237–285.

Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. 2024. ParaHome: Parameterizing Everyday Home Activities Towards 3D Generative Modeling of Human-Object Interactions. *arXiv preprint arXiv:2401.10232* (2024).

Lucas Kovar and Michael Gleicher. 2003. Flexible automatic motion blending with registration curves.. In *Symposium on Computer Animation*, Vol. 2. San Diego, CA, USA.

Lucas Kovar and Michael Gleicher. 2004. Automated extraction and parameterization of motions in large data sets. *ACM Transactions on Graphics (ToG)* 23, 3 (2004), 559–568.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion graphs. *ACM Transactions on Graphics* 21, 3 (2002), 473–482.

Jehee Lee, Jinxiang Chai, Paul SA Reitsma, Jessica K Hodgins, and Nancy S Pollard. 2002. Interactive control of avatars animated with human motion data. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 491–500.

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. 2023a. Controllable human-object interaction synthesis. *arXiv preprint arXiv:2312.03913* (2023).

Jiaman Li, Alexander Clegg, Roozbeh Mottaghi, Jiajun Wu, Xavier Puig, and C Karen Liu. 2025. Controllable human-object interaction synthesis. In *European Conference on Computer Vision*. Springer, 54–72.

Jiaman Li, Jiajun Wu, and C Karen Liu. 2023b. Object motion guided human motion synthesis. *ACM Transactions on Graphics (TOG)* 42, 6 (2023), 1–11.

Libin Liu and Jessica Hodgins. 2018. Learning basketball dribbling skills using trajectory optimization and deep reinforcement learning. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Libin Liu, KangKang Yin, Michiel van de Panne, Tianjia Shao, and Weiwei Xu. 2010. Sampling-based contact-rich motion control. *ACM Transactions on Graphics (TOG)* 29, 4 (2010), 1–10.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 2022. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21013–21022.

Yun Liu, Bowen Yang, Licheng Zhong, He Wang, and Li Yi. 2024. Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. *arXiv preprint arXiv:2412.17730* (2024).

Zhengyi Luo, Jinkun Cao, Sammy Christen, Alexander Winkler, Kris M Kitani, and Weipeng Xu. 2024. Omnigrasp: Grasping Diverse Objects with Simulated Humanoids. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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.

Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. 2021. Isaac Gym: High Performance GPU Based Physics Simulation For Robot Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. https://openreview.net/forum?id=fgFBtYgJQX_

Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. 2023. Mimicgen: A data generation system for scalable robot learning using human demonstrations. *arXiv preprint arXiv:2310.17596* (2023).

Matteo Menolotto, Dimitrios-Sokratis Komaris, Salvatore Tedesco, Brendan O'Flynn, and Michael Walsh. 2020. Motion capture technology in industrial applications: A

systematic review. *Sensors* 20, 19 (2020), 5687.

Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. 2024. Synthesizing physically plausible human motions in 3d scenes. In *2024 International Conference on 3D Vision (3DV)*. IEEE, 1498–1507.

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.

Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel van de Panne. 2018. DeepMimic. *ACM Transactions on Graphics* (Aug 2018), 1–14. https://doi.org/10.1145/3197517.3201311

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergexuey Levine, and Sanja Fidler. 2022. ASE: Large-scale Reusable Adversarial Skill Embeddings for Physically Simulated Characters. *ACM Trans. Graph.* 41, 4, Article 94 (July 2022).

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. 2021. AMP: Adversarial Motion Priors for Stylized Physics-Based Character Control. *ACM Transactions on Graphics* (Aug 2021), 1–20. https://doi.org/10.1145/3450626.3459670

Wilbert Pumacay, Ishika Singh, Jiafei Duan, Ranjay Krishna, Jesse Thomason, and Dieter Fox. 2024. The colosseum: A benchmark for evaluating generalization for robotic manipulation. *arXiv preprint arXiv:2402.08191* (2024).

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

Carmelo Sferrazza, Dun-Ming Huang, Xingyu Lin, Youngwoon Lee, and Pieter Abbeel. 2024. HumanoidBench: Simulated Humanoid Benchmark for Whole-Body Locomotion and Manipulation. *arXiv preprint arXiv:2403.10506* (2024).

Sebastian Starke, Yiwei Zhao, Taku Komura, and Kazi Zaman. 2020. Local motion phases for learning multi-contact character movements. *ACM Trans. Graph.* 39, 4 (aug 2020). https://doi.org/10.1145/3386569.3392450

Sebastian Starke, Yiwei Zhao, Fabio Zinno, and Taku Komura. 2021. Neural animation layering for synthesizing martial arts movements. *ACM Trans. Graph.* 40, 4 (jul 2021). https://doi.org/10.1145/3450626.3459881

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. 2020. GRAB: A Dataset of Whole-Body Human Grasping of Objects. In *European Conference on Computer Vision (ECCV)*. https://grab.is.tue.mpg.de

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

Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. 2023. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–9.

Jiashun Wang, Jessica Hodgins, and Jungdam Won. 2024b. Strategy and skill learning for physics-based table tennis animation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. 2023b. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 11359–11366.

Yinhuai Wang, Jing Lin, Ailing Zeng, Zhengyi Luo, Jian Zhang, and Lei Zhang. 2023a. Physhoi: Physics-based imitation of dynamic human-object interaction. *arXiv preprint arXiv:2312.04393* (2023).

Yinhuai Wang, Qihan Zhao, Runyi Yu, Ailing Zeng, Jing Lin, Zhengyi Luo, Hok Wai Tsui, Jiwen Yu, Xiu Li, Qifeng Chen, et al. 2024c. Skillmimic: Learning reusable basketball skills from demonstrations. *arXiv preprint arXiv:2408.15270* (2024).

Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. 2024a. Move as You Say Interact as You Can: Language-guided Human Motion Generation with Scene Affordance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 433–444.

Jungdam Won and Jehee Lee. 2019. Learning body shape variation in physics-based characters. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–12.

Zeqi Xiao, Tai Wang, Jingbo Wang, Jinkun Cao, Wenwei Zhang, Bo Dai, Dahua Lin, and Jiangmiao Pang. 2024. Unified Human-Scene Interaction via Prompted Chain-of-Contacts. In *The Twelfth International Conference on Learning Representations*.

Pei Xu and Ruocheng Wang. 2024. Synchronize Dual Hands for Physics-Based Dexterous Guitar Playing. In *SIGGRAPH Asia 2024 Conference Papers*. 1–11.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. 2023a. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14928–14940.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. 2024. InterDreamer: Zero-Shot Text to 3D Dynamic Human-Object Interaction. *arXiv preprint arXiv:2403.19652* (2024).

Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruicheng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, et al. 2023b. Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4737–4746.

Jie Yang, Xuesong Niu, Nan Jiang, Ruimao Zhang, and Siyuan Huang. 2025. F-HOI: Toward Fine-grained Semantic-Aligned 3D Human-Object Interactions. In *European Conference on Computer Vision*. Springer, 91–110.

Chengwen Zhang, Yun Liu, Ruofan Xing, Bingda Tang, and Li Yi. 2024b. Core4d: A 4d human-object-human interaction dataset for collaborative object rearrangement. *arXiv preprint arXiv:2406.19353* (2024).

Chong Zhang, Wenli Xiao, Tairan He, and Guanya Shi. 2024c. WoCoCo: Learning Whole-Body Humanoid Control with Sequential Contacts. *arXiv preprint arXiv:2406.06005* (2024).

Hui Zhang, Sammy Christen, Zicong Fan, Otmar Hilliges, and Jie Song. 2025. Graspxl: Generating grasping motions for diverse objects at scale. In *European Conference on Computer Vision*. Springer, 386–403.

Haotian Zhang, Ye Yuan, Viktor Makoviychuk, Yunrong Guo, Sanja Fidler, Xue Bin Peng, and Kayvon Fatahalian. 2023b. Learning Physically Simulated Tennis Skills from Broadcast Videos. *ACM Trans. Graph.* (2023), 14 pages. https://doi.org/10.1145/3592408

Xiaoyu Zhang, Matthew Chang, Pranav Kumar, and Saurabh Gupta. 2024a. Diffusion Meets DAgger: Supercharging Eye-in-hand Imitation Learning. *arXiv preprint arXiv:2402.17768* (2024).

Yunbo Zhang, Deepak Gopinath, Yuting Ye, Jessica Hodgins, Greg Turk, and Jungdam Won. 2023a. Simulation and retargeting of complex multi-character interactions. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–11.

Liming Zhao and Alla Safonova. 2009. Achieving good connectivity in motion graphs. *Graphical Models* 71, 4 (2009), 139–152.

A ADDITIONAL EXPERIMENT

A.1 Evaluation on Data Efficiency

To evaluate our method's improvement in data efficiency, we conduct experiments with varying amounts of training data for a single skill. Specifically, we construct four datasets from BallPlay-M's pickup clips with increasing sizes: 1, 4, 10, and 40 clips respectively. For each dataset, we train policies using both SkillMimic (SM) and our method (SM+STF+ATS+HE) for approximately 3.2 billion samples. During evaluation, we place balls randomly within concentric circles of varying radii (1-5 meters) around the humanoid. The quantitative results in Tab. 4 show our method consistently outperforms baselines across all data scales. Even with 40 training clips, it improves generalization success rate by 13% (reaching 96%), demonstrating its effectiveness scales well with increased training data.

A.2 Evaluation on Locomotion Skills

Although our method is primarily designed for learning interaction skills, we investigate its potential benefits for locomotion skill generalization. We selected two representative locomotion skills from the BallPlay-M dataset: a single-clip Run skill and a Getup skill comprising eight diverse Getup clips. These skills were trained simultaneously using a unified policy with skill conditions.

For comparison, we also evaluated against state-of-the-art locomotion methods, specifically AMP [Peng et al. 2021] combined with random state initialization [Peng et al. 2022], denoted as AMP-RSI. Other baseline settings follow those in the BallPlay-M experiment in the main paper, except that object-related terms were removed from both observations and reward functions since this experiment focuses on pure locomotion.

Tab. 5 presents the quantitative results, demonstrating the effectiveness of our approach in enhancing robustness and generalization performance of locomotion skills.

A.3 Evaluation on Zero-Shot In-Hand Reorientation

While existing methods excel at grasp pose generation [Jiang et al. 2021; Luo et al. 2024; Wang et al. 2023b; Xu et al. 2023b; Zhang et al. 2025], they typically cannot generate complex in-hand manipulations. Our method can effectively bridges this gap by augmenting discrete grasp frames into continuous manipulations. Specifically, to reorientate a cube to target poses, we first obtain a grasp pose using existing methods [Zhang et al. 2025]. Given the geometric symmetry of the cube under 90-degree rotations, we can augment a single grasp pose into 24 valid grasp poses (6 faces × 4 orientations). Each pose is then replicated for 100 frames to create 24 trajectories, with the cube's 3D orientation serving as the condition label c. We train a pose-conditioned policy using our full method. During testing, given a novel cube orientation as condition, our method successfully generates natural hand manipulation sequences to achieve the desired cube orientation. Fig. 1(d) in the main paper shows an example of 90-degree cube orientation, where the intermediate manipulation process is learned from no demonstration. This application demonstrates our method's potential for both manipulation learning and data augmentation.

Table 4. Performance under different data amounts of ball pickup.

| Method | SR↑ (%) with Random Ball Positions |
	1 Clip	4 Clips	10 Clips	40 Clips
SM	0.10	16.26	32.26	82.84
SM + Ours	**0.54**	**46.64**	**85.68**	**96.32**

Table 5. Quantitative comparison on locomotion skills.

| Method | \multicolumn{2}{c}{SR↑ (%) / εNSR↑ (%) / NR ↑} | | \multicolumn{2}{c}{TSR↑ (%)} |
	Getup	Run	Getup-Run	Run-Getup
AMP + RSI	**99.3** / **98.6** / 0.01	93.3 / 80.5 / 0.66	37.9	99.7
DM	24.4 / 24.9 / 0.64	46.5 / 22.2 / 0.73	0.2	22.4
DM + ε-NSI	96.9 / 97.9 / 0.47	93.2 / 84.9 / 0.65	62.6	5.7
DM + Ours	98.5 / 98.2 / 0.54	97.1 / 81.5 / 0.64	67.2	96.2
SM	69.2 / 66.0 / **0.80**	74.0 / 36.4 / **0.78**	10.5	44.9
SM + ε-NSI	97.8 / 97.3 / 0.66	99.4 / **91.8** / 0.77	93.4	12.7
SM + Ours	99.1 / 98.0 / 0.64	**99.9** / 91.0 / 0.71	**97.9**	**100.0**

B TECHNICAL DETAILS

B.1 Observation

The state $s_t = \{o_t^{sbj}, o_t^f, o_t^{obj}\}$ observed by the policy consists of the following components:

- Humanoid observation o_t^{sbj}:
  - Global root height
  - Body position and rotation in local coordinates
  - Body position velocity and angular velocity
- Contact observation o_t^f:
  - Net contact forces at fingertips
- Object observation o_t^{obj}:
  - Position and rotation in local coordinates
  - Linear and angular velocities

All coordinates are transformed into the humanoid's root local coordinate system to enhance generalization.

B.2 Policy

The policy output is parameterized as a Gaussian distribution:

$$a_t \sim \mathcal{N}(\phi_\pi(s_t, h_t, c), \Sigma_\pi), \tag{11}$$

where ϕ_π is a three-layer MLP (1024-512-512 units, ReLU activations) that maps state s_t, history embedding h_t, and skill condition c to action means. The variances Σ_π are set to 0.055 during training for exploration and 0 during testing for stability. The action a_t represents target joint rotations, which are processed by a PD controller to generate joint torques.

B.3 Reward Function

Following SkillMimic [Wang et al. 2024c], we use a unified imitation reward for RLID training. The imitation reward combines four components:

$$r_t = S(s_{t+1}, \hat{s}_{t+1}) = r_t^b * r_t^o * r_t^{rel} * r_t^{cg}, \tag{12}$$

where r_t^b, r_t^o, r_t^{rel}, and r_t^{cg} represent body motion, object motion, relative motion, and contact graph rewards respectively. All reward weights are listed in Tab. 8.

- **Body Motion Term:**

$$r_t^b = r_t^p * r_t^r * r_t^{pv} * r_t^{rv}, \tag{13}$$

Each sub-term follows:

$$r_t^\alpha = e^{-\lambda^\alpha * \text{MSE}(s_{t+1}^\alpha, \hat{s}_{t+1}^\alpha)}, \tag{14}$$

where $\alpha \in \{p, r, pv, rv\}$ represents position, rotation, position velocity, and rotation velocity respectively. $\hat{s}_{t+1}^\alpha$ and s_{t+1}^α denote reference and simulated states.

- **Object Motion Term:**

$$r_t^o = r_t^{op} * r_t^{or} * r_t^{opv} * r_t^{orv}, \tag{15}$$

with sub-terms following the same formulation as body motion rewards.

- **Relative Motion Term:**

$$r_t^{rel} = e^{-\lambda^{rel} * \text{MSE}(s_{t+1}^{rel}, \hat{s}_{t+1}^{rel})}, \tag{16}$$

- **Contact Graph Term:**

$$r_t^{cg} = e^{-\sum_{j=1}^J \lambda^{cg}[j] * e_{t+1}^{cg}[j]}, \tag{17}$$

where $e_{t+1}^{cg} = |s_{t+1}^{cg} - \hat{s}_{t+1}^{cg}|$ represents the contact error between simulated and reference states. J is the number of contact pairs. For experiment on BallPlay-M, s^{cg} contains three contact pairs: ball-hands contact, ball-body contact, and body-hands contact. Due to Isaac Gym's limitations in detecting complex contact pairs [Makoviychuk et al. 2021], we determine contacts based on contact forces. For experiment on ParaHome [Kim et al. 2024], the contact graph reward is disabled (i.e., $r_t^{cg} = 1$).

For DM, we follow the implementation of SM, with the only modification being the adoption of DM-style additive reward:

$$r_t = r_t^p + r_t^r + r_t^{rv} + r_t^{op}, \tag{18}$$

with reward weights listed in Tab. 9.

B.4 Connection Rules

Given any two states s_A and s_B, their kinematic similarity is evaluated using a modified similarity metric that excludes contact information:

$$S_k(s_A, s_B) = r^b * r^o * r^{rel}, \tag{19}$$

where r^b, r^o, and r^{rel} are defined identically to those in the reward function. Let $\beta = S_k(s_A, s_B)$ denote the computed similarity score, the connection from s_A to s_B is established according to the following criteria, where τ represents a predetermined similarity threshold:

- When $\beta > \tau$, we introduce intermediate masked states between s_A and s_B. The number of masked states is determined by:

$$N = \min(-\lfloor \log_{10}(\beta) \rfloor, N_{max}), \tag{20}$$

where N_{max} denotes the maximum allowable number of masked states.

Table 6. Hyperparameters for policy training.

Parameter	Value
dim(c) Skill Embedding Dimension	64
Σ_π Action Distribution Variance	0.055
Samples Per Update Iteration	65536
Policy/Value Function Minibatch Size	16384
γ Discount	0.99
Adam Stepsize	2×10^{-5}
GAE(λ)	0.95
TD(λ)	0.95
PPO Clip Threshold	0.2
T Episode Length	60
μ Dimension of history embedding	3
k History horizon length	60

Table 7. Hyperparameters for Data Augmentation. Note that N_{max} represents the max allowable number of masked states; τ means the state similarity threshold; p_e is the probability of sampling from external reference states; p_n is the probability of sampling from external states; λ_s is the Adaptive Trajectory Sampling (ATS) weight; λ_c is the inter-class ATS weight.

Parameter	BallPlay-M	Locomotion	ParaHome
N_{max}	10	10	10
τ	1×10^{-10}	1×10^{-10}	1×10^{-10}
p_e	0.1	0.1	–
p_n	0.1	0.1	0.1
λ_s	10	10	10
λ_c	5	5	5
$\varepsilon_{rootpos}$	0.1	0.1 for Run. 1.0 for Getup	0.1
$\varepsilon_{rootvel}$	0.1	0.1 for Run. 1.0 for Getup	0.1
$\varepsilon_{rootrot}$	0.1	0.1	0.1
$\varepsilon_{rootrotvel}$	0.1	0.1	0.1
ε_{dof}	0.1	0.1	0.1
ε_{dofvel}	0.1	0.1	0.1
ε_{objpos}	0.1	0.1	0.1
$\varepsilon_{objposvel}$	0.1	0.1	0.1
ε_{objrot}	0.1	0.1	0.1
$\varepsilon_{objrotvel}$	0.1	0.1	0.1

- When $\beta < \tau$, the connection is deemed invalid and subsequently discarded from consideration.

For BallPlay-M [Wang et al. 2024c], when constructing the stitched trajectory graph (STG), we apply coordinate transformation to align the stitched state pairs. Specifically, for a state s_A, we first transform its root position by aligning its (x,y) coordinates with the reference state s_B's root before computing their similarity. The transformed s_A is then evaluated against connection criteria to determine whether it should be added to the STG.

B.5 History Encoder Pre-training

We present a self-supervised pre-training approach for the History Encoder (HE) that enables effective learning of temporal dependencies. The pre-training process follows a behavioral cloning paradigm, where we jointly train an encoder θ to generate compact historical embeddings and an state predictor ψ to model state transitions. The encoder θ consists of three convolutional layers followed by a fully

---

ALGORITHM 1: Online Motion Data Augmentation

Input: Dataset $\mathcal{D}$, probabilities p_1, p_2, neighborhood radius ϵ
Output: Augmented motion sequence $\tilde{\mathbf{m}}$
Sample reference skill motion $\mathbf{m} = \{\hat{s}_0, ..., \hat{s}_T\}$ from $\mathcal{D}$;
if *Bernoulli(p_1)* **then**
  Sample reference skill motion $\mathbf{n} \neq \mathbf{m}$ from $\mathcal{D}$;
  Sample initial state $\hat{s}^*$ from $\mathbf{n}$;
end
else
  Sample $k \in [0, T]$ according to Eq. 8;
  $\hat{s}^* \leftarrow \hat{s}_k$;
end
if *Bernoulli(p_2)* **then**
  Sample neighborhood states $\hat{s}_{nb} \sim \mathcal{N}(\hat{s}^*, \epsilon)$;
  $\hat{s}^* \leftarrow \hat{s}_{nb}$;
end
for $i \in [0, T]$ **do**
  Compute similarity scores $d_i = \text{Dist}(\hat{s}^*, \hat{s}_i)$ (Eq. 19);
end
Find closest state index $j = \arg\min_i d_i$;
Calculate mask length N according to d_j following Sec. B.4;
Return augmented sequence $\tilde{\mathbf{m}} = \{\hat{s}^*, \underbrace{s_\varnothing, ..., s_\varnothing}_{N}, \hat{s}_j, ..., \hat{s}_T\}$;

Table 8. Reward weights of SM in different tasks.

Parameter	BallPlay-M	Locomotion	ParaHome
λ^p Position	20	20	20
λ^r Rotation	20	20	20
λ^{pv} Velocity	0	0	0
λ^{rv} Rotation Velocity	0	0	0
λ^{op} Object Position	1	–	1
λ^{or} Object Rotation	0	–	0
λ^{opv} Object Velocity	0	–	0
λ^{orv} Object Angular Velocity	0	–	0
λ^{rel} Relative Motion	20	–	20
$\boldsymbol{\lambda}^{cg}[0]$ Ball-Hands Contact	5	–	–
$\boldsymbol{\lambda}^{cg}[1]$ Ball-Body Contact	5	–	–
$\boldsymbol{\lambda}^{cg}[2]$ Body-Hands Contact	5	–	–

Table 9. Reward weights of DM in different tasks. DM related settings are tested only on Locomotion and BallPlay tasks.

Parameter	BallPlay-M	Locomotion
λ^p Position	40	40
λ^r Rotation	2	2
λ^{rv} Rotation Velocity	0.1	0.1
λ^{op} Object Position	40	–

connected layer, while the predictor ψ employs a three-layer MLP structure similar to the policy network.

Given a demonstration dataset, we randomly sample state trajectories $\{\hat{s}_{t-k}, ..., \hat{s}_{t+1}\}$ as reference sequences with their corresponding condition labels c. The History Encoder θ processes a sequence of k historical states to generate a μ-dimensional embedding $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \theta(s_{t-k}, ..., s_{t-1}) \tag{21}$$

This historical context is concatenated with the current state $\hat{s}_t$ and condition c, then passed to a state transition predictor ψ, which estimates the next state:

$$s_{t+1} = \psi([c, \hat{s}_t, \boldsymbol{h}_t]) \tag{22}$$

The training objective combines state prediction accuracy with an embedding regularization term:

$$\mathcal{L} = \lambda_a|\hat{s}_{t+1} - s_{t+1}|^2 + \lambda_b|\boldsymbol{h}_t|^2 \tag{23}$$

where $\lambda_a = 1$ and $\lambda_b = 10^{-5}$ are hyperparameters controlling the balance between prediction accuracy and embedding regularization. Both θ and ψ are optimized during pre-training through this objective. The predictor ψ effectively approximates the combined behavior of the policy and physical simulator, ensuring that successful convergence during pre-training indicates the HE has learned meaningful temporal representations.

C HYPERPARAMETERS

The hyperparameters for policy training are detailed in Tab. 6. Additionally, Tab. 7 presents the data augmentation hyperparameters, Tab. 8 and Tab. 9 displays the reward weights for SM and DM respectively.