ComplexMimic: Human–Scene Interaction Imitation in Complex 3D Environments

Lu Pan[1,2]* and Hongwei Zhao[1,2]

[1] College of Computer Science and Technology, Jilin University, China
[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, China
panlutag@gmail.com, zhaohw@jlu.edu.cn

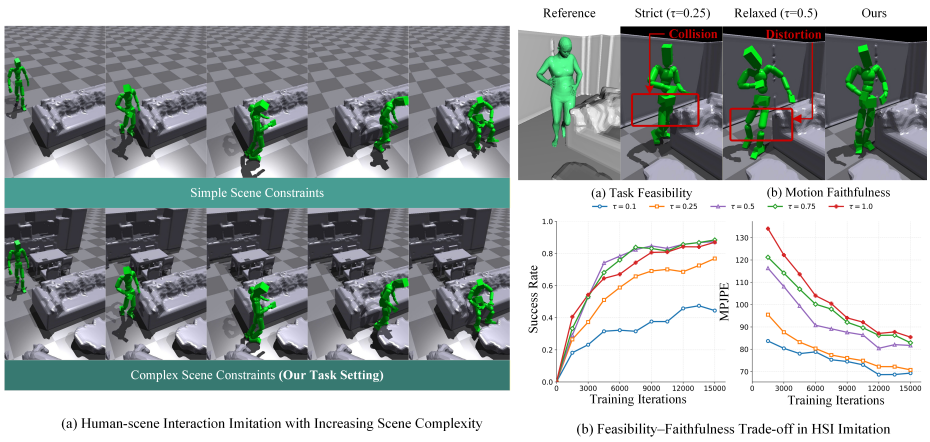
Abstract. Physics-based Human-Scene Interaction (HSI) imitation learning is crucial for embodied intelligence as it bridges the gap between kinematic 3D motions and real-world dynamics. However, most existing methods focus on simplified scene settings, leaving complex environments largely unexplored, which limits their applicability in real-world scenarios. In this paper, we focus on HSI mimicry in complex environments. Under this complex setting, we observe an inherent trade-off between successfully performing interaction and maintaining natural, physically plausible motions. To address this challenge, we propose ComplexMimic, a framework that reconstructs diverse HSI by interpreting imperfect MoCap data. First, we introduce a Dual Flow Strategy, which learns two complementary experts: an imitation expert for accurate motion tracking and an interaction expert for collision-aware adaptation in complex scenes. Second, naive multi-expert distillation, which treats all experts equally, often under-samples challenging behaviors, limiting effective learning. To mitigate this issue, we propose a difficulty-aware distillation strategy that adaptively weights supervision and prioritizes hard-yet-learnable trajectories guided by failure statistics and learning progress signals. Extensive experiments on three benchmark datasets demonstrate that our approach outperforms current state-of-the-art methods. Our implementation is available at `https://github.com/LuPan23/ComplexMimic`.

Keywords: Humanoid Imitation Learning, Human-Scene Interaction, Difficulty-Aware Distillation

1 Introduction

Humanoid whole-body control via imitation learning plays a crucial role in embodied intelligence and physically plausible character animation [8, 26, 27, 39, 44, 45]. In recent years, many studies have made significant progress on multiple tasks, including single-human motion imitation [8, 19, 21, 35, 39], human-object motion imitation [21, 26–28, 35], human-scene interaction [19], and multi-human interaction [20].

* Corresponding author



(a) Human-scene Interaction Imitation with Increasing Scene Complexity

(b) Feasibility–Faithfulness Trade-off in HSI Imitation

Fig. 1: (a) Prior humanoid imitation-learning studies [19] typically focus on scene-free or overly simplified environments (upper row), whereas our work targets realistic HSI imitation in complex 3D scenes (lower row). (b) Feasibility–faithfulness trade-off in HSI imitation. Varying the early termination threshold τ during training reveals a clear trade-off: larger τ improves task feasibility (higher success rate) but reduces motion fidelity (higher MPJPE), while smaller τ enforces faithful motion at the cost of task success.

Humans can easily move through cluttered spaces. Even in scenes filled with furniture, narrow gaps, and irregular geometry, we naturally avoid collisions while maintaining the intent of our motion. Replicating this capability for physically simulated humanoids remains challenging, especially for the imitation of Human–Scene Interaction (HSI) within complex environments. This task requires tracking the given reference trajectory as faithfully as possible while respecting the geometric constraints imposed by complex scene meshes. However, the existing HSI imitation learning approach focuses only on overly simplified scenes and limited HSI behavior. For instance, MimicBench [19] typically contains a single object as scene configurations and includes only six predefined tasks, such as sitting on a chair or lying on a sofa (Fig. 1).

In this paper, we focus on how to mimic human motions, captured from MoCap data, that dynamically interact with complex scenes. As illustrated in Fig. 1(a), a humanoid is expected to imitate reference motions while avoiding collisions with complex scene geometry. Through our investigation, we empirically observe an important pattern. As illustrated in Fig. 1(b), the early termination threshold τ, which controls the maximum allowed deviation from the reference motion during training, has a significant impact on both the success rate and MPJPE of the learned policy. A larger τ improves task feasibility by allowing greater flexibility to avoid collisions, but degrades motion fidelity. Conversely, a smaller τ imposes stricter motion tracking but limits the policy's ability to explore feasible motions, leading to premature failures in complex scenes.

Motivated by this challenge, we propose a two-stage method, **ComplexMimic**, to address the trade-off between motion fidelity and task feasibility. In stage I, we present the Dual Flow Strategy that learns two specialized expert policies. Specifically, we first train an *imitation expert* in a scene-free setting using a strict termination threshold $\tau \leq 0.25$, emphasizing precise motion tracking and high motion fidelity. Then, we learn another *interaction expert* that leverages the full scene mesh and additional scene-aware information (*i.e.*, height map). With a relaxed termination threshold, *i.e.*, $\tau \geq 0.5$, this expert enables robust adaptation to environmental geometry and physical collision handling.

In stage II, we distill the complementary expertise of the two experts into a single unified policy via multi-teacher knowledge distillation [10, 14]. However, naive distillation strategies using uniform or random sampling overlook differences in expert learning difficulty, which may overtrain on easy samples, and consequently exhibit low efficiency. To tackle this problem, we introduce **D**ifficulty-**A**ware **D**istillation (DAD), which adaptively balances supervision across experts and prioritizes hard yet informative samples. We partition the training motions into two groups, *i.e.*, imitation group and interaction group, so that each expert can specialize in the motions it handles best. The motion difficulty score is computed from online failure statistics, which guide inter-group sampling (favoring more difficult groups) and intra-group prioritization of harder samples.

We evaluate ComplexMimic on large-scale HSI datasets, including TRUMANS [16] and LINGO [15]. Experimental results demonstrate that our approach achieves substantially better feasibility–faithfulness balance compared to existing baselines, especially achieving a **3.90%** increase in Succ and a **9.36%** reduction in $E_{\text{g-mpjpe}}$ on the TRUMANS dataset. Furthermore, zero-shot transfer to the real-world scanned dataset GIMO [51] demonstrates strong robustness and generalization. Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work systematically addressing human-scene interaction imitation under such realistic and complicated scene constraints. We propose a two-stage framework, ComplexMimic, to tackle this challenging problem.
- We reveal a feasibility–faithfulness trade-off, where relaxing the early termination tracking constraint improves interaction feasibility but degrades motion fidelity, and vice versa.
- We propose a Dual-Flow Strategy, which trains an imitation expert with strict tracking constraint for motion faithfulness and an interaction expert with relaxed tracking constraint and scene observations for feasible scene interaction.
- We introduce Difficulty-Aware Distillation, a multi-teacher distillation strategy that adaptively balances supervision across experts and prioritizes hard-yet-learnable motion trajectories based on online motion-wise failure statistics and learning progress signals.

2 Related Work

2.1 Humanoid Imitation Learning

Physics-based imitation learning has become an efficient and practical paradigm for training humanoid controllers [8, 9, 17, 19, 38, 39, 43–45]. Early work, such as DeepMimic [26], demonstrates that reinforcement learning guided by tailored imitation rewards can reproduce highly dynamic motions with stable physical control. Subsequent efforts relaxed strict motion alignment requirements. AMP [28] introduces a novel discriminator-based objective, allowing more flexible matching between simulated motion and data. ASE [27] further improves motion diversity by learning reusable low-level skills that can be composed for downstream tasks. More recently, approaches that focus on scaling control capacity and improving tracking robustness have been introduced. PHC [21] adopts a Mixture-of-Experts (MoE) [14] paradigm and a negative sample mining strategy to enhance training efficiency. MaskedMimic [35] leverages a transformer-based full-body controller to achieve better tracking performance and utilizes a masked motion modeling to achieve more flexible humanoid robot control.

Although promising results have been made in humanoid imitation learning, these advances are mainly focused on simplified environments without scene constraints or with limited scene settings [19, 21, 26, 35]. In contrast, we target physics-based human–scene interaction imitation under dense and complicated 3D environments, where geometric constraints introduce frequent collision and instability. We validate our framework on recent large-scale human–scene interaction datasets, including TRUMANS [16] and LINGO [15], which provide substantially more challenging and realistic interaction scenarios.

2.2 Human-Scene Interaction

Human-Scene Interaction (HSI) is a long-standing problem in human motion modeling [2,3,36,46–48] and animation synthesis [1,4,5,7,12,13,18,23,31,37,40, 49,50]. A major line of work in HSI focuses on kinematic modeling, which models body-scene spatial relationships to generate motions consistent with scene geometry without requiring a closed-loop physics controller. PROX [5] is an early and influential work on this line, which proposes a novel human-scene interaction dataset based on RGB sequences in 12 scenes with 20 subjects. Building on PROX [5], POSA [6] introduces a body-centric contact-and-semantics representation over SMPL-X [25] vertices and is trained on PROX to extend to unseen scenes with plausible contact and semantic consistency. Recent advances focus on end-to-end generative models. HUMANISE [41] and LINGO [15] tend to generate 3D human motions in scenes based on text prompts through diffusion models [11, 32–34], while TRUMANS [16] experiments with an autoregressive diffusion model conditioned on action labels. Another line of research adopts a closed-loop physics controller. UNIHSI [42] leverages large language models and a chain of contact to achieve human-scene interaction control. TokenHSI [24] proposes a transformer-based unified controller for human-scene interaction control and skill learning.

Although prior works have achieved progress in physics-based human-scene interaction control, full-body imitation learning for HSI with complex 3D environments is less explored. In contrast, we target physics-based HSI imitation learning, where a humanoid must reproduce a given reference motion as faithfully as possible under scene constraints. And we explicitly address the fidelity-feasibility conflict by decoupling motion-faithful tracking and scene-aware interaction into complementary experts, and reconciling them through adaptive distillation with hierarchical difficulty-aware sampling.

3 Methodology

Task Formulation. The goal of human-scene interaction (HSI) imitation learning is to learn a policy π that produces a simulated human-scene interaction motion sequence $\{q_t\}_{t=1}^{T}$, which closely matches a ground-truth reference $\{\hat{q}_t\}_{t=1}^{T}$ derived from the MoCap data. Given the physical simulation of the humanoid and the 3D scene constraints, the policy π should also compensate for missing or inaccurate details in the dataset. In particular, each simulated motion q_t is defined as $q_t = \{\theta_t, p_t\}$, where $\theta_t \in \mathbb{R}^{J\times 3}$ and $p_t \in \mathbb{R}^{J\times 3}$ represent the rotations and positions for J joints, respectively. All simulation states have the corresponding ground-truth values, denoted by the hat symbol.

Overview. We formulate HSI imitation as a Markov Decision Process (MDP) $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, defined by states $\mathcal{S}$, actions $\mathcal{A}$, simulator-provided transition dynamics P, reward function r, and discount factor γ. Fig. 2 illustrates our two-stage training framework. In Stage I, a Dual Flow Strategy trains decoupled teacher experts $\pi^{(T)}$: an imitation expert $\pi^{(T)}_{\text{Imitate}}$ for motion tracking accuracy and an interaction expert $\pi^{(T)}_{\text{Interact}}$ for interaction feasibility. In Stage II, we adaptively distill the knowledge of these teachers into a scalable student policy $\pi^{(S)}$ for whole-body control via Difficulty-Aware Distillation (DAD). Sec. 3.1 details how teacher policies (*i.e.*, experts) are trained with reinforcement learning, and Sec. 3.2 presents DAD for multi-teacher balancing and efficient mining of hard-yet-learnable motion samples.

3.1 Dual Flow Strategy

We propose a Dual Flow Strategy to train two teacher policies specialized for motion tracking and interaction, respectively.

Imitation Expert. Given an initialized motion state s_t, the imitation expert $\pi^{(T)}_{\text{Imitate}}$ predicts an action a_t, which is executed in the physics simulator $\mathcal{E}(\mathcal{S})$ to obtain the next state s_{t+1} and compute the imitation reward r_t. With a horizon length of L_H, the motion episode terminates when it satisfies an early termination condition or reaches the horizon length. After executing all motions in a minibatch, the policy is optimized using the PPO objective [30]:

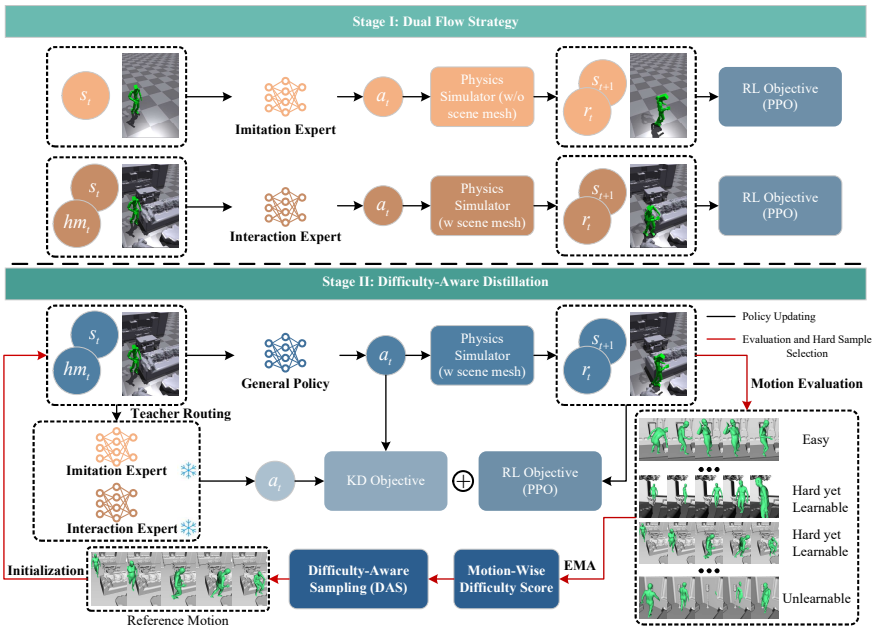


Fig. 2: Overview of ComplexMimic, a two-stage framework for human–scene interaction imitation in complex environments. In Stage I, an imitation expert and an interaction expert are trained to achieve motion-faithful tracking and collision-aware adaptation, respectively. In Stage II, these experts are frozen and distilled into a unified policy via Difficulty-Aware Distillation (DAD). Black arrows denote policy updates; red arrows represent evaluation and hard-sample selection.

$\max_{\theta_I} \mathbb{E}\left[\sum_t \gamma^t r^{\text{imit}}(s_t, m_t)\right]$, where r^{imit} denotes the imitation reward and m_t represents the reference motion. Detailed definition of r^{imit} is given in the supplementary material.

Interaction Expert. The interaction expert $\pi^{(T)}_{\text{Interact}}(a_t|s_t, hm_t)$ is trained in the scene-constrained simulator $\mathcal{E}(\mathcal{S})$. It receives a height map [35] of the surrounding 3D environment hm_t as additional scene perception. Specifically, the height map is represented by 20×20 samples on a square grid centered on the character and aligned with its heading direction, with neighboring samples spaced by 8 cm. The policy, therefore, conditions on both the humanoid state s_t and the scene representation hm_t to produce collision-aware actions while still tracking the reference motion using the imitation reward r^{imit}.

Policy Learning. Following [28], the control policy π is optimized using PPO [30]. The policy gradient objective is defined as

$$\mathcal{L}_{\text{PPO}}(\theta) = \mathbb{E}_t\Big[\min\big(r_t(\theta)A_t,\ \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)A_t\big)\Big], \tag{1}$$

where θ denotes the parameters of the policy π, and $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the likelihood ratio between the updated and old policies. ϵ is a small clipping constant that constrains policy updates, and A_t is the advantage estimate computed using the generalized advantage estimator (GAE(λ)) [29].

3.2 Difficulty-Aware Distillation

As illustrated in Fig. 2, we utilize **D**ifficulty-**A**ware **D**istillation (DAD) to adaptively balance decoupled experts and prioritize hard yet learnable motions. Specifically, DAD consists of three components: (i) regime-specialized teacher routing, which assigns each motion to the most suitable expert, (ii) a multi-teacher distillation objective that transfers complementary knowledge to the student policy, and (iii) difficulty-aware sampling that prioritizes informative motions during training. Details of each component are given in this section.

Regime-Specialized Teacher Routing A naive multi-teacher distillation strategy typically combines supervision from multiple experts via action averaging or random routing, which often introduces gradient interference because different experts are optimized for fundamentally different objectives. As a result, the student may learn behaviors that are neither fully faithful to the reference motion nor sufficiently feasible under scene constraints.

To address this issue, we adopt regime-specialized teacher routing, which assigns each motion to the expert that handles it most reliably. We partition the training motions into two groups according to the tracking performance of the imitation expert in the scene-aware simulator: the tracking-friendly group $\mathcal{G}_{\text{Imitate}}$ and the interaction-critical group $\mathcal{G}_{\text{Inter}}$. Regime-specialized teacher routing is defined as

$$\pi^{(T)}(m) = \begin{cases} \pi^{(T)}_{\text{Imitate}}, & m \in \mathcal{G}_{\text{Imitate}}, \\ \pi^{(T)}_{\text{Interact}}, & m \in \mathcal{G}_{\text{Inter}}. \end{cases} \tag{2}$$

The groups are constructed by evaluating $\pi^{(T)}_{\text{Imitate}}$ on all training motions before distillation: successfully tracked motions are assigned to $\mathcal{G}_{\text{Imitate}}$, and the rest to $\mathcal{G}_{\text{Inter}}$. The effectiveness of regime-specialized routing is validated in Sec. 4.4.

Multi-Teacher Distillation Objective. Under the regime-specialized teacher routing, for each student transition (s_t, a_t, r_t), we query the teacher actions $a_t^{\text{imitate}} \sim \pi^{(T)}_{\text{Imitate}}(\cdot \mid s_t)$ and $a_t^{\text{inter}} \sim \pi^{(T)}_{\text{Interact}}(\cdot \mid s_t, hm_t)$. Assuming Gaussian policies $\pi(a|s) = \mathcal{N}(\mu(s), \Sigma(s))$, distillation is implemented via the Kullback–Leibler divergence D_{KL} between action distributions:

$$\begin{aligned} \mathcal{L}_{\text{KD}}(\theta) = \mathbb{I}[m \in \mathcal{G}_{\text{Inter}}]\; D_{\text{KL}}\Big(\pi_\theta(\cdot|s_t, hm_t) \;\Big\|\; \pi^{(T)}_{\text{Interact}}(\cdot|s_t, hm_t)\Big) \\ + \Big(1 - \mathbb{I}[m \in \mathcal{G}_{\text{Inter}}]\Big)\, D_{\text{KL}}\Big(\pi_\theta(\cdot|s_t, hm_t) \;\Big\|\; \pi^{(T)}_{\text{Imitate}}(\cdot|s_t)\Big). \end{aligned} \tag{3}$$

Distillation alone is insufficient because the student must interact with the environment under its own state distribution in $\mathcal{E}(\mathcal{S})$. Therefore, the student is trained with a joint objective:

$$\mathcal{L}(\theta) = \lambda_{\text{PPO}}\,\mathcal{L}_{\text{PPO}}(\theta) + \lambda_{\text{KD}}\,\mathcal{L}_{\text{KD}}(\theta), \tag{4}$$

where $\mathcal{L}_{\text{PPO}}$ denotes the standard clipped PPO objective computed from rollouts of π_θ in $\mathcal{E}(\mathcal{S})$, and $\mathcal{L}_{\text{KD}}$ is the distillation loss. The coefficients λ_{PPO} and λ_{KD} balance reinforcement learning and distillation.

Difficulty-Aware Sampling. Uniform or random sampling is a common choice in multi-teacher distillation with imitation learning [39, 44], implicitly assuming that all motions contribute equally. In complex 3D scenes, this assumption breaks down: difficulty is highly skewed, and a small subset of motions accounts for most failures. Uniform sampling, therefore, wastes computation on already-solved trajectories while under-training feasibility-critical ones. To address this challenge, we propose the **D**ifficulty-**A**ware **S**ampling (DAS), which reallocates training budget based on the student's online difficulty and learnability signals.

We first define the motion-wise difficulty score with online failure statistics. Let $m \in \mathcal{M}$ denote a motion clip. Every K iterations, we evaluate the current student on motion m and compute a binary failure indicator $f_k(m) = \mathbb{I}[\text{failure on } m \text{ at checkpoint } k]$, where "failure" is defined as violating the tracking-success criterion, namely when the average joint distance to the reference exceeds 0.5 m at any timestep. We maintain a smoothed motion difficulty score via EMA as $D_k(m) \leftarrow (1-\beta)D_{k-1}(m) + \beta f_k(m)$, so $D_k(m) \in [0,1]$ measures how persistently motion m fails under the current student.

Then, we utilize the motion-wise difficulty score to achieve inter-group sampling. To avoid collapsing into easy regimes, we perform inter-group sampling based on the motion difficulty score. We partition motions into aforementioned C groups $\{\mathcal{G}_c\}_{c=1}^{C}$. The class difficulty is defined by aggregating motion difficulties:

$$D_k^{\text{cls}}(c) = \frac{1}{|\mathcal{G}_c|} \sum_{m \in \mathcal{G}_c} D_k(m), \tag{5}$$

and we sample a group according to a softmax distribution:

$$p_k(c) = \frac{\exp(\tau_{\text{inter}}\, D_k^{\text{cls}}(c))}{\sum_{c'=1}^{C} \exp(\tau_{\text{inter}}\, D_k^{\text{cls}}(c'))}, \tag{6}$$

where τ_{inter} controls how strongly we focus on difficult groups.

We further prioritize informative motion samples via intra-group prioritization. After selecting a group $c \sim p_k(c)$, we perform intra-group prioritization to mine the most informative motions within the group. We define a per-motion mining score $s_k(m)$ and sample motions using:

$$p_k(m \mid c) = \frac{\exp(\tau_{\text{intra}}\, s_k(m))}{\sum_{m' \in \mathcal{G}_c} \exp(\tau_{\text{intra}}\, s_k(m'))}, \tag{7}$$

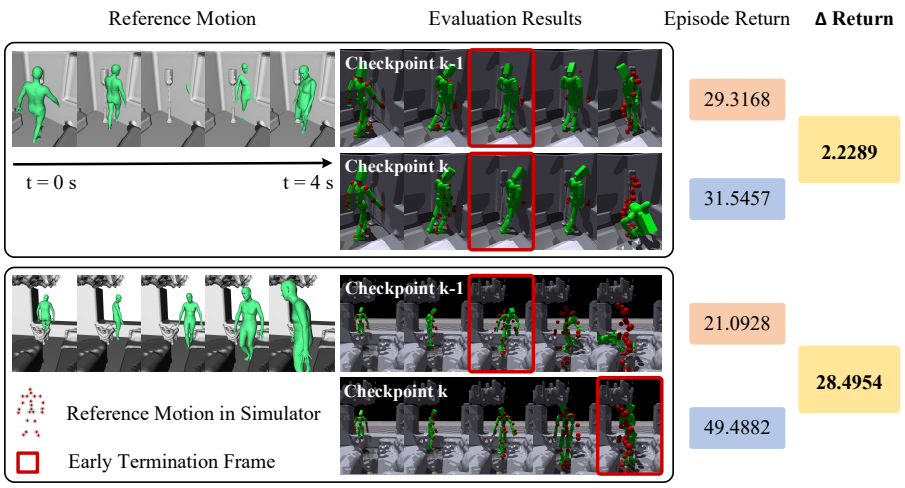


Fig. 3: Examples of return improvement under different motion cases. The first reference motion conflicts heavily with the scene meshes, resulting in limited return gains. In contrast, the second motion achieves better tracking performance and larger return gains, demonstrating the effectiveness of our learning progress signal.

A simple and effective choice is to set $s_k(m)$ proportional to the motion difficulty, $s_k(m) = D_k(m)$. Therefore, motions that fail more frequently are mined more aggressively within each group.

Naïve intra-group hard mining may over-focus on motions that remain stagnant for long periods (e.g., highly disturbed or effectively infeasible cases), wasting computation and destabilizing training. We therefore introduce a learnability signal based on return improvement. Let $R_k(m)$ be the student's episode return on motion m at checkpoint k. We define the return improvement as $\Delta R_k(m) = R_k(m) - R_{k-1}(m)$, and maintain its exponential moving average (EMA) as $I_k(m) \leftarrow (1-\eta)I_{k-1}(m) + \eta\,\Delta R_k(m)$.

As shown in Fig. 3, some motions exhibit consistent return improvements, while others remain stagnant. If improvement is below a small threshold ϵ_u, we downweight the sampling probability with:

$$u_k(m) = \exp\Big(-\lambda_u \max(0, \epsilon_u - I_k(m))\Big), \tag{8}$$

where λ_u controls the strength of penalization. We then incorporate this term into intra-group prioritization by redefining the mining score as $s_k(m) = D_k(m) + \log u_k(m)$. This keeps mining focused on motions that are hard yet learnable, while suppressing stagnant outliers.

Putting everything together, the final probability of sampling motion m is

$$p_k(m) = (1-\epsilon)\, p_k(c(m))\, p_k(m \mid c(m)) + \epsilon \cdot \frac{1}{|\mathcal{M}|}, \tag{9}$$

where $c(m)$ maps motion m to its group, and ϵ controls the strength of a small uniform prior over motions.

4 Experiments

4.1 Experiments Setup

Datasets. All the policies are trained on the TRUMANS dataset [16], which is the largest high-quality human-scene interaction dataset with diverse and challenging scene geometries. Following SceneMI [13], motion sequences are segmented into 121 frames, with 10% randomly selected for the test set. For evaluation, we tested our approach on three settings: (i) in-domain evaluation on TRUMANS [16] dataset; (ii) out-of-domain evaluation on unseen scene conditions on LINGO [15] dataset; and (iii) real-world evaluation on GIMO [51] dataset, which consists of scanning real-world scene meshes with reconstruction artifacts and noises.

Metrics. Following PHC [21], we utilize a series of pose-based and physics-based metrics to evaluate our motion imitation performance. We report the success rate (Succ), deeming imitation unsuccessful when, at any point during imitation, the body joints are on average > 0.5m from the reference motion. The global mean per-joint position error (MPJPE) $E_{\text{g-mpjpe}}$ and root-relative MPJPE E_{mpjpe} are presented to measure the imitator's ability of imitating the reference motion both globally and locally (root-relative). To show physical realism, we also compare acceleration E_{acc} (mm/frame^2) and velocity E_{vel} (mm/frame) difference between simulated and MoCap motion. Detailed definitions of these metrics are provided in the supplementary material.

Implementation Details. All models are trained on a single NVIDIA 4090 GPU. Physics simulation is carried out in NVIDIA's Isaac Gym [22]. The control policy is run at 30 Hz, while the simulation runs at 60 Hz. Following [21], we do not consider body shape variation and use the mean SMPL body shape for evaluation. The temperatures for inter-group sampling and intra-group prioritization are both fixed to 1.0. Uniform mixture ϵ is set at a ratio 0.05. EMA of Failure and return statistics are set to 0.95 and 0.9, respectively. The return improvement threshold ϵ_u is 0.1, and λ_u is chosen as 1.0. Expert policies are implemented as MLPs with hidden sizes $\{2048, 1024, 1024, 512\}$. The student policy is implemented as a four-layer Transformer encoder with 4 heads, a hidden size of 512, and a feed-forward layer of 2048. The critic model is implemented as an MLP with hidden sizes $\{1024, 512\}$. We use $\tau = 0.25$ and $\tau = 0.5$ for the imitation expert and interaction expert, respectively. For student training, we keep τ fixed at 0.5 to encourage physically feasible rollouts. Additional implementation details are provided in the supplementary material.

Table 1: Quantitative comparisons on the TRUMANS [16] dataset. Best results are in **bold** and second-best are underlined.

Method	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
DeepMimic [26]	0.674	99.002	73.349	16.335	14.823
AMP [28]	0.802	94.670	76.022	17.366	14.122
PHC [21]	0.853	85.793	65.365	13.559	12.055
MaskedMimic [35]	0.872	120.260	105.099	26.913	24.717
Ours w/o DAS	0.878	84.026	64.213	**12.759**	11.194
Ours	**0.906**	**77.840**	**64.128**	12.931	**10.721**

Table 2: Quantitative comparisons on the LINGO [15] dataset.

Method	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
DeepMimic [26]	0.535	128.974	92.421	24.079	20.907
AMP [28]	0.635	124.200	93.327	25.679	20.652
PHC [21]	0.615	**120.472**	**84.782**	23.648	20.383
MaskedMimic [35]	0.719	163.525	132.315	38.387	36.499
Ours w/o DAS	0.678	122.046	87.072	21.643	18.595
Ours	**0.746**	124.665	90.575	**20.641**	**17.303**

Baseline Models. We compare ComplexMimic with four state-of-the-art physics-based motion imitation methods: DeepMimic [26], AMP [28], PHC [21], and MaskedMimic [35]. For fair comparison, we use the official implementations and augment all baselines with height maps at the same scale as ours.

4.2 Quantitative Evaluation

As shown in Tab. 1, our method consistently outperforms the competitive approaches on the TRUMANS [16] test set across all metrics. In particular, Succ improves from 0.872 to 0.906 compared to the strongest baseline MaskedMimic [35], while simultaneously reducing both $E_{\text{g-mpjpe}}$ and E_{mpjpe}. This demonstrates that ComplexMimic effectively resolves the feasibility–fidelity trade-off in complicated scene settings. Tab. 2 presents out-of-domain evaluation results on the LINGO dataset [15] with unseen scene configurations. Our method achieves the highest Succ while maintaining competitive tracking accuracy compared to PHC [21], indicating strong cross-scene generalization. Tab. 3 further reports real-world evaluation results on the GIMO dataset [51], which contains scanned environments with reconstruction noise. Our approach consistently surpasses the baseline methods across all metrics, achieving the best overall performance. These results highlight the robustness and generalization capability of our dual flow

Table 3: Quantitative comparisons on the GIMO [51] dataset.

Method	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
DeepMimic [26]	0.365	178.929	114.317	46.385	38.374
AMP [28]	0.446	173.632	113.682	46.974	37.353
PHC [21]	0.443	173.580	109.696	47.151	38.556
MaskedMimic [35]	0.405	286.629	191.210	69.821	68.460
Ours w/o DAS	0.515	<u>165.117</u>	**102.009**	<u>38.738</u>	<u>31.643</u>
Ours	**0.579**	**161.969**	<u>104.590</u>	**36.845**	**29.749**

Table 4: Ablation study on TRUMANS [16]. "w/o" indicates removal; "–" indicates replacement of regime-specialized teacher routing. Additional ablation studies on LINGO [15] and GIMO [51] are provided in the supplementary material.

Method	Succ ↑	$E_{\text{g-mpjpe}}$ ↓	E_{mpjpe} ↓	E_{acc} ↓	E_{vel} ↓
w/o Imitation Expert	0.876	86.059	64.487	13.170	11.477
w/o Interaction Expert	0.847	82.489	<u>62.464</u>	**12.565**	11.416
w/o Inter-Group Sampling	<u>0.899</u>	82.045	62.816	<u>12.779</u>	<u>10.918</u>
w/o Intra-Group Prioritization	0.898	<u>80.735</u>	**61.016**	13.349	11.248
w/o Learnability Filtering	0.877	88.497	67.465	14.231	11.876
w/o PPO	0.830	110.880	79.460	18.788	16.475
- Random Routing	0.874	83.016	63.715	12.625	10.945
- Action Averaging	0.873	83.698	64.075	12.712	11.030
Ours	**0.906**	**77.840**	64.128	12.931	**10.721**

strategy and difficulty-aware distillation framework. We further analyze the contribution of individual components in Sec. 4.4.

4.3 Qualitative Evaluation

To further demonstrate the advantage of our method against other competing baselines, Fig. 4 presents visual examples on the TRUMANS [16] test set. AMP [28] often fails to robustly track the reference in complicated scenes, leading to noticeable penetrations and unstable contacts. While PHC [21] can accurately track the reference motion, it is prone to being disturbed by the 3D scene meshes. As for the MaskedMimic [35], although it can achieve feasible HSI tracking, the deviations from the reference are rather obvious. In contrast, our approach produces more feasible interactions by correcting incorrect body configurations while preserving the overall motion structure.

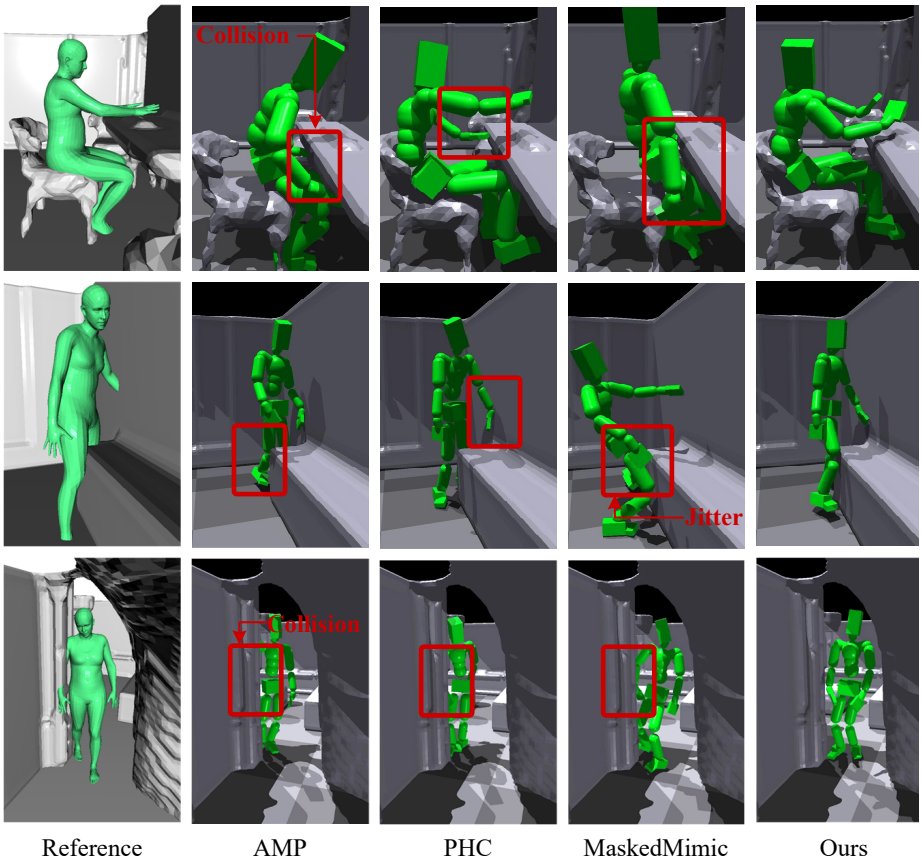


Fig. 4: Qualitative comparisons on the TRUMANS [16] dataset. Red boxes highlight typical failure modes under complicated scenes, such as undesired collisions, unstable contacts, and noticeable deviations from the reference. More visual results can be found in the supplementary material.

4.4 Ablation Study

Effectiveness of Dual Flow Strategy. As shown in Tab. 4, removing either teacher degrades the performance significantly. Without the imitation expert, tracking error $E_{\text{g-mpjpe}}$ degrades (from 77.840 to 86.059), while removing the interaction expert leads to a notable drop in Succ (from 0.906 to 0.847), highlighting the importance of interaction-aware supervision. Qualitative results in Fig. 5 further confirm that the two experts provide complementary guidance for fidelity and feasibility.

Effectiveness of Difficulty-Aware Sampling (DAS). Each component of DAS consistently contributes to the final performance. Disabling intra-group prioritization or inter-group sampling reduces Succ and worsens tracking accuracy,

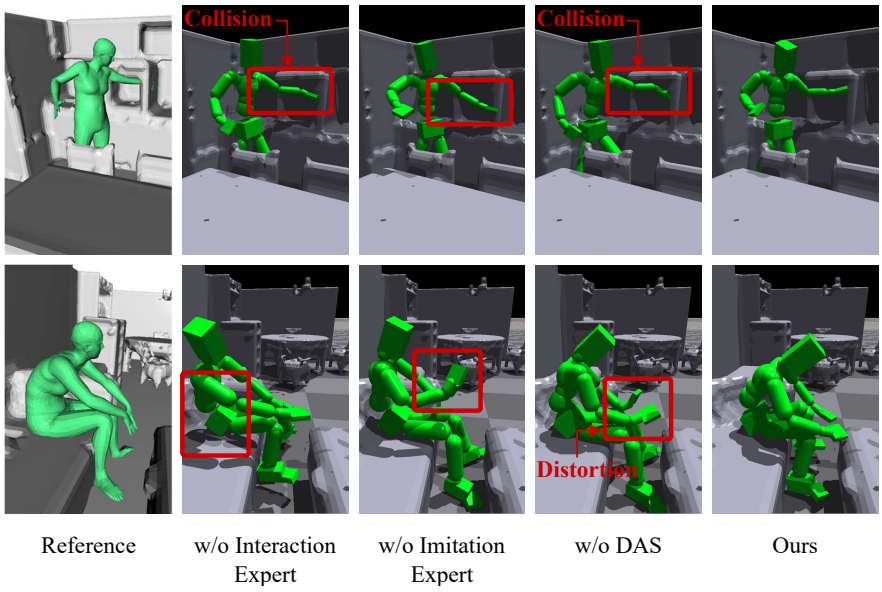


Fig. 5: Qualitative ablation comparisons on TRUMANS [16].

showing that (i) focusing updates on harder motions within each regime and (ii) balancing training across regimes are both beneficial. The learnability filtering is particularly important: removing it significantly degrades both Succ (from 0.906 to 0.877) and $E_{\text{g-mpjpe}}$ (from 77.840 to 88.497), suggesting that suppressing hard but unlearnable clips is necessary for stable and efficient distillation.

Effectiveness of joint optimization with PPO. As exhibited in Tab. 4, adding the PPO objective significantly improves both Succ (from 0.830 to 0.906) and $E_{\text{g-mpjpe}}$ (from 110.880 to 77.840), highlighting the necessity of on-policy refinement beyond pure distillation.

Effectiveness of regime-specialized teacher routing. To evaluate the necessity of regime-specialized teacher routing, we compare with two alternatives: (i) Random routing, where each motion is supervised by either expert with equal probability, and (ii) Action averaging, where the student imitates the mean action of both experts. As shown in Tab. 4, both variants degrade performance. Random routing introduces inconsistent supervision across motion regimes, while action averaging blends conflicting expert objectives, leading to gradient interference and suboptimal performance.

5 Conclusion

In this paper, we address an important yet underexplored problem: human-scene interaction imitation learning under complex scene constraints. To solve the scene-induced infeasibility and achieve accurate motion imitation, we propose a decoupled training paradigm, which leverages the imitation expert to track the reference motion faithfully and utilizes the interaction expert to interact with the 3D environment appropriately. To combine these complementary capabilities into a unified policy, we introduce difficulty-aware distillation, which adaptively balances supervision across different expert regimes and prioritizes hard-yet-learnable motions using failure statistics and return improvement signals. Extensive experiments on three human-scene interaction datasets show that our approach consistently improves both interaction feasibility and motion accuracy, and validate the contribution of each component.

References

1. Collorone, L., Gioia, M., Pappa, M., Leoni, P., Ficarra, G., Litany, O., Spinelli, I., Galasso, F.: Monster: a unified model for motion, scene, text retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10940–10949 (2025)
2. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5152–5161 (2022)
3. Guo, C., Zuo, X., Wang, S., Cheng, L.: Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In: European Conference on Computer Vision. pp. 580–597. Springer (2022)
4. Hassan, M., Ceylan, D., Villegas, R., Saito, J., Yang, J., Zhou, Y., Black, M.J.: Stochastic scene-aware motion prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11374–11384 (2021)
5. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2282–2292 (2019)
6. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14708–14718 (2021)
7. Hassan, M., Guo, Y., Wang, T., Black, M., Fidler, S., Peng, X.B.: Synthesizing physical character-scene interactions. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–9 (2023)
8. He, T., Gao, J., Xiao, W., Zhang, Y., Wang, Z., Wang, J., Luo, Z., He, G., Sobanbab, N., Pan, C., et al.: Asap: Aligning simulation and real-world physics for learning agile humanoid whole-body skills. arXiv preprint arXiv:2502.01143 (2025)
9. He, T., Luo, Z., He, X., Xiao, W., Zhang, C., Zhang, W., Kitani, K., Liu, C., Shi, G.: Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. arXiv preprint arXiv:2406.08858 (2024)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)

11. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
12. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13274–13285 (2022)
13. Hwang, I., Zhou, B., Kim, Y.M., Wang, J., Guo, C.: Scenemi: Motion in-betweening for modeling human-scene interaction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6034–6045 (2025)
14. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
15. Jiang, N., He, Z., Wang, Z., Li, H., Chen, Y., Huang, S., Zhu, Y.: Autonomous character-scene interaction synthesis from text instruction. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
16. Jiang, N., Zhang, Z., Li, H., Ma, X., Wang, Z., Chen, Y., Liu, T., Zhu, Y., Huang, S.: Scaling up dynamic human-scene interaction modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1737–1747 (2024)
17. Liao, Q., Truong, T.E., Huang, X., Gao, Y., Tevet, G., Sreenath, K., Liu, C.K.: Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion. arXiv preprint arXiv:2508.08241 (2025)
18. Liu, X., Hou, H., Yang, Y., Li, Y.L., Lu, C.: Revisit human-scene interaction via space occupancy. In: European Conference on Computer Vision. pp. 1–19. Springer (2024)
19. Liu, Y., Yang, B., Zhong, L., Wang, H., Yi, L.: Mimicking-bench: A benchmark for generalizable humanoid-scene interaction learning via human mimicking. arXiv preprint arXiv:2412.17730 (2024)
20. Liu, Z., Ge, J., Xiong, M., Gu, J., Tang, B., Jing, W., Chen, S.: It takes two: Learning interactive whole-body control between humanoid robots (2025)
21. Luo, Z., Cao, J., Kitani, K., Xu, W., et al.: Perpetual humanoid control for real-time simulated avatars. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10895–10904 (2023)
22. Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., et al.: Isaac gym: High performance gpu-based physics simulation for robot learning. arXiv preprint arXiv:2108.10470 (2021)
23. Pan, L., Wang, J., Huang, B., Zhang, J., Wang, H., Tang, X., Wang, Y.: Synthesizing physically plausible human motions in 3d scenes. In: 2024 International Conference on 3D Vision (3DV). pp. 1498–1507. IEEE (2024)
24. Pan, L., Yang, Z., Dou, Z., Wang, W., Huang, B., Dai, B., Komura, T., Wang, J.: Tokenhsi: Unified synthesis of physical human-scene interactions through task tokenization. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5379–5391 (2025)
25. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10975–10985 (2019)
26. Peng, X.B., Abbeel, P., Levine, S., Van de Panne, M.: Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. ACM Transactions On Graphics (TOG) **37**(4), 1–14 (2018)

27. Peng, X.B., Guo, Y., Halper, L., Levine, S., Fidler, S.: Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. ACM Transactions On Graphics (TOG) **41**(4), 1–17 (2022)
28. Peng, X.B., Ma, Z., Abbeel, P., Levine, S., Kanazawa, A.: Amp: Adversarial motion priors for stylized physics-based character control. ACM Transactions on Graphics (ToG) **40**(4), 1–20 (2021)
29. Schulman, J., Moritz, P., Levine, S., Jordan, M., Abbeel, P.: High-dimensional continuous control using generalized advantage estimation. arXiv preprint arXiv:1506.02438 (2015)
30. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
31. Shen, Y., Liu, H., Zhang, L., Liu, P., Xia, R., Yao, T., Feng, T.: Detach: Cross-domain learning for long-horizon tasks via mixture of disentangled experts. arXiv preprint arXiv:2508.07842 (2025)
32. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: nternational conference on machine learning. pp. 2256–2265 (2015)
33. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
34. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
35. Tessler, C., Guo, Y., Nabati, O., Chechik, G., Peng, X.B.: Maskedmimic: Unified physics-based character control through masked motion inpainting. ACM Transactions On Graphics (TOG) **43**(6), 1–21 (2024)
36. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. arXiv preprint arXiv:2209.14916 (2022)
37. Wang, W., Pan, L., Dou, Z., Mei, J., Liao, Z., Lou, Y., Wu, Y., Yang, L., Wang, J., Komura, T.: Sims: Simulating stylized human-scene interactions with retrieval-augmented script generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14117–14127 (2025)
38. Wang, Y., Zhao, Q., Yu, R., Tsui, H.W., Zeng, A., Lin, J., Luo, Z., Yu, J., Li, X., Chen, Q., et al.: Skillmimic: Learning basketball interaction skills from demonstrations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17540–17549 (2025)
39. Wang, Y., Yang, M., Ding, Z., Zhang, Y., Zeng, W., Xu, X., Jiang, H., Lu, Z.: From experts to a generalist: Toward general whole-body control for humanoid robots. arXiv preprint arXiv:2506.12779 (2025)
40. Wang, Z., Chen, Y., Jia, B., Li, P., Zhang, J., Zhang, J., Liu, T., Zhu, Y., Liang, W., Huang, S.: 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. pp. 433–444 (2024)
41. Wang, Z., Chen, Y., Liu, T., Zhu, Y., Liang, W., Huang, S.: Humanise: Language-conditioned human motion generation in 3d scenes. Advances in Neural Information Processing Systems **35**, 14959–14971 (2022)
42. Xiao, Z., Wang, T., Wang, J., Cao, J., Zhang, W., Dai, B., Lin, D., Pang, J.: Unified human-scene interaction via prompted chain-of-contacts. arXiv preprint arXiv:2309.07918 (2023)
43. Xu, P., Wu, Z., Wang, R., Sarukkai, V., Fatahalian, K., Karamouzas, I., Zordan, V., Liu, C.K.: Learning to ball: Composing policies for long-horizon basketball moves. ACM Transactions on Graphics (TOG) **44**(6), 1–14 (2025)

44. Xu, S., Ling, H.Y., Wang, Y.X., Gui, L.Y.: Intermimic: Towards universal whole-body control for physics-based human-object interactions. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 12266–12277 (2025)
45. Yu, R., Wang, Y., Zhao, Q., Tsui, H.W., Wang, J., Tan, P., Chen, Q.: Skillmimic-v2: Learning robust and generalizable interaction skills from sparse and noisy demonstrations. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)
46. Zhang, J., Fan, H., Yang, Y.: Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17592–17602 (2025)
47. Zhang, J., Zhang, Y., Cun, X., Zhang, Y., Zhao, H., Lu, H., Shen, X., Shan, Y.: Generating human motion from textual descriptions with discrete representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14730–14740 (2023)
48. Zhang, M., Cai, Z., Pan, L., Hong, F., Guo, X., Yang, L., Liu, Z.: Motiondiffuse: Text-driven human motion generation with diffusion model. IEEE transactions on pattern analysis and machine intelligence **46**(6), 4115–4128 (2024)
49. Zhao, K., Wang, S., Zhang, Y., Beeler, T., Tang, S.: Compositional human-scene interaction synthesis with semantic control. In: European Conference on Computer Vision (ECCV) (2022)
50. Zhao, K., Zhang, Y., Wang, S., Beeler, T., Tang, S.: Synthesizing diverse human motions in 3d indoor scenes. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14738–14749 (2023)
51. Zheng, Y., Yang, Y., Mo, K., Li, J., Yu, T., Liu, Y., Liu, C.K., Guibas, L.J.: Gimo: Gaze-informed human motion prediction in context. In: European Conference on Computer Vision. pp. 676–694. Springer (2022)