BRIC: Bridging Kinematic Plans and Physical Control at Test Time

Dohun Lim[1], **Minji Kim**[1], **Jaewoon Lim**[1], **Sungchan Kim**[1*]

[1]Department of Computer Science and Artificial Intelligence, Jeonbuk National University, Korea
{dohun.lim, kmkmj927, ljy3513, s.k}@jbnu.ac.kr

Abstract

We propose BRIC, a novel test-time adaptation (TTA) framework that enables long-term human motion generation by resolving execution discrepancies between diffusion-based kinematic motion planners and reinforcement learning-based physics controllers. While diffusion models can generate diverse and expressive motions conditioned on text and scene context, they often produce physically implausible outputs, leading to execution drift during simulation. To address this, BRIC dynamically adapts the physics controller to noisy motion plans at test time, while preserving pre-trained skills via a loss function that mitigates catastrophic forgetting. In addition, BRIC introduces a lightweight test-time guidance mechanism that steers the diffusion model in the signal space without updating its parameters. By combining both adaptation strategies, BRIC ensures consistent and physically plausible long-term executions across diverse environments in an effective and efficient manner. We validate the effectiveness of BRIC on a variety of long-term tasks, including motion composition, obstacle avoidance, and human-scene interaction, achieving state-of-the-art performance across all tasks.

1 Introduction

Despite recent advances in computer vision, graphics, and robotics that have enabled substantial progress in generating realistic human motions from text and scene inputs (Xu et al. 2023; Yang et al. 2024; Diller and Dai 2024; Song et al. 2024; Xiao et al. 2024; Pan et al. 2024; Tessler et al. 2024; Wu et al. 2025b; Chen et al. 2025; Xu et al. 2025b,a), the generation of *physically plausible long-term motion* remains an open challenge. Diffusion-based text-to-motion models (Tevet et al. 2023; Zhang et al. 2024) have recently gained significant attention by leveraging large-scale datasets such as HumanML3D (Guo et al. 2022). These models require the composition of diverse motion primitives into coherent long-term sequences (Barquero, Escalera, and Palmero 2024) to achieve smooth transitions (Barquero, Escalera, and Palmero 2024; Zhao, Li, and Tang 2025), reliable obstacle avoidance (Karunratanakul et al. 2023), and fine-grained human-scene interaction (HSI) (Xiao et al. 2024; Chen et al. 2025).

Despite their expressiveness, diffusion-based kinematic models often fail to enforce physical constraints, leading to unrealistic artifacts such as foot skating and floating. To address these shortcomings, recent approaches adopt a two-stage pipeline that improves physical realism by aligning motion sequences executed by a reinforcement learning (RL)-based controller with noisy plans generated by a diffusion model, typically by adapting the planner to the policy (Serifi et al. 2024) or vice versa (Ren et al. 2023; Tevet et al. 2025; Xu et al. 2025a). However, this two-stage structure introduces several limitations.

First, the inherent noise in the generated plans makes physics-based controllers prone to error accumulation over time, leading to drift that impedes the execution of long-term motion sequences (Tevet et al. 2025; Xu et al. 2025a). Second, tuning the policy to track noisy motion plans, as done in recent works (Xu et al. 2025a; Tevet et al. 2025), can induce catastrophic forgetting, resulting in overfitting to specific styles or tasks and hindering generalization. Third, diffusion-based models typically cannot perform scene-aware planning or obstacle avoidance without additional supervision (Yi et al. 2024; Caesar et al. 2020). To avoid retraining these motion planners, test-time guidance has emerged as an alternative means of controlling diffusion-based outputs to satisfy such user-defined constraints (Huang et al. 2024; Rempe et al. 2023; Janner et al. 2022; Karunratanakul et al. 2024, 2023). However, since this method does not enforce physical constraints, the guided motions may still exhibit self-collisions or physically implausible contacts.

To address these challenges, we propose *BRIC*, a test-time adaptation (TTA) framework (Wang et al. 2021, 2022) that tightly couples diffusion-based kinematic planners with physics-based controllers. BRIC features two key components: (1) a TTA-based motion policy that robustly tracks noisy motion plans, and (2) a lightweight signal-space test-time guidance strategy that aligns motion plans with the controller's executed motion distribution without requiring costly backpropagation through the diffusion model.

BRIC adapts a physics-based controller (the source domain) to the noisy motion plans generated by a diffusion model (the target domain) by updating the policy parameters online in response to distributional shifts or input noise. Crucially, our approach preserves previously acquired skills via a catastrophic forgetting-aware loss (Ebrahimi et al. 2020;

*Corresponding author

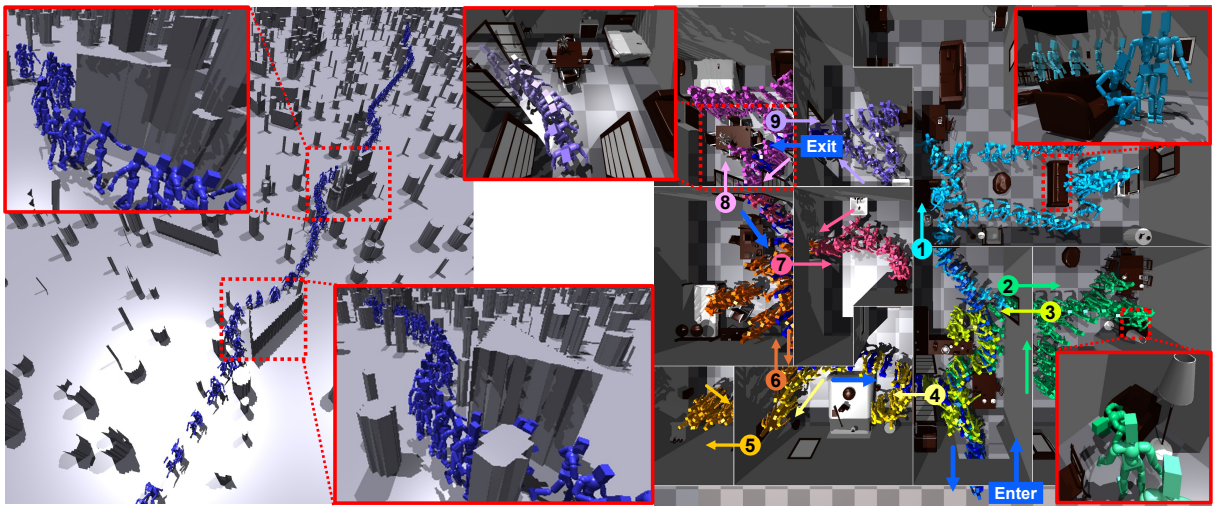

Figure 1: *BRIC* introduces a test-time adaptation (TTA) framework for *physically plausible long-term motion* generation, effectively and efficiently bridging the execution gap between autoregressive diffusion-based kinematic planners and physics-based controllers. It enables robust execution of highly extended and challenging tasks, such as (left) reliable *obstacle avoidance over 100* m and (right) navigation through an indoor environment spanning about *12 minutes for human-scene interaction*, with the visit order of rooms annotated in the figure. Animated videos of these tasks, along with additional visualizations, are available on the project page at https://bric2026.github.io/.

Parisi et al. 2019). While conventional test-time guidance methods (Huang et al. 2024; Karunratanakul et al. 2023; Janner et al. 2022) generate motions that satisfy user-defined objectives, they often rely on repeated backpropagation through the diffusion model, resulting in significant computational overhead. To overcome this limitation, we introduce an efficient test-time guidance strategy that rapidly generates motion plans within the proposed TTA framework without backpropagation, enabling fast and scalable inference.

By integrating these strategies, BRIC produces motions that are both consistent with the intended plans and physically executable. We evaluate BRIC on long-term tasks including text-driven motion composition, obstacle avoidance, and HSI, demonstrating state-of-the-art performance across all benchmarks. Our main contributions are as follows:

- We introduce BRIC, a novel TTA framework that adapts a physics-based controller to noisy motion plans generated by a diffusion-based kinematic planner.
- We propose a new loss function that mitigates catastrophic forgetting, allowing the motion policy to adapt while preserving essential skills.
- We develop a computationally efficient test-time guidance method that avoids backpropagation through the diffusion model and generalizes to unseen environments.
- We demonstrate that BRIC achieves state-of-the-art results on challenging long-term tasks, including motion composition, obstacle avoidance, and HSI.

2 Related Work

Test-Time Adaptation in Generative and Control Models. Test-time adaptation (TTA) refers to techniques that dynamically update model parameters during deployment to accommodate distribution shifts. TENT (Wang et al. 2021) addresses this by minimizing prediction entropy to update batch normalization parameters, while CoTTA (Wang et al. 2022) mitigates catastrophic forgetting using exponential moving average predictions and stochastic resets. More recently, TTA has been extended to generative models. For instance, Diffusion-TTA (Prabhudesai et al. 2023) adapts a classifier to test-time inputs based on the outputs from a diffusion model, improving robustness under out-of-distribution conditions (Tsai et al. 2024; Prabhudesai et al. 2023; Yu et al. 2023). BRIC extends this paradigm to RL by adapting a physics-based controller to the noisy motion plan distribution generated by a diffusion model. Unlike CoTTA, which focuses on classification, BRIC addresses the more complex problem of adapting an autoregressive control policy for physics-based state transitions, introducing unique challenges specific to control.

Test-Time Guidance for Diffusion-Based Motion Generation. Diffusion models (Ho, Jain, and Abbeel 2020) have shown strong performance in both image (Yin et al. 2024) and motion (Tevet et al. 2023; Zhang et al. 2024) generation, often employing classifier-free guidance (Ho and Salimans 2021) to incorporate conditioning signals at inference time. Recent work has explored this paradigm by guiding diffusion-based motion generation through the optimization of objective functions during sampling. Such objectives include RL value functions (Rempe et al. 2023), spatial waypoints (Janner et al. 2022), keyframes (Karunratanakul et al. 2023, 2024), obstacle avoidance (Rempe et al. 2023; Karunratanakul et al. 2024), and scene-level constraints (Zhao, Li, and Tang 2025; Yi et al. 2024). These methods typically operate in the latent space, converting samples to the signal space and optimizing them based on task-specific objectives. However, they

require gradient computation through the diffusion model, which incurs substantial computational cost and limits their applicability in TTA settings. Moreover, the resulting guided motions are often physically implausible when deployed in physics simulations, leading to execution failures. To overcome these limitations, BRIC introduces an efficient test-time guidance method that directly optimizes task objectives in the signal space, without relying on backpropagation through the diffusion model.

Bridging Diffusion Models and Physics-Based Controllers. An increasing body of research aims to integrate kinematic diffusion models (Tevet et al. 2023; Zhang et al. 2024) with physics-based controllers (Luo et al. 2023) to generate physically plausible motion (Yuan et al. 2023; Ren et al. 2023; Wu et al. 2025a; Serifi et al. 2024; Tevet et al. 2025). PhysDiff (Yuan et al. 2023) reduces unrealistic artifacts by projecting generated motions into simulation, while InsActor (Ren et al. 2023) trains a RL controller using a differentiable simulator. CLoSD (Tevet et al. 2025) establishes a closed loop between a diffusion-based planner and an RL controller in an autoregressive manner, whereas RobotMDM (Serifi et al. 2024) aligns the diffusion planner offline using the controller's value function. UniPhys (Wu et al. 2025a) unifies the planner and controller into a single architecture, but at the cost of modularity and composability. PARC (Xu et al. 2025a) alternates training between the motion planner and the RL policy by using each other's outputs as self-supervised signals. However, PARC trains both components under a shared data distribution, whereas our method explicitly handles distribution shifts at deployment time.

3 Method

3.1 Preliminaries

Diffusion Based Motion Planning. We adopt the autoregressive diffusion model DiP from CLoSD (Tevet et al. 2025) as our motion planner, denoted by G, which generates a motion sequence $\boldsymbol{x}_0^{1:H} \in \mathbb{R}^{H\times d}$, where H is the sequence length and d is the dimensionality of each motion frame. The forward process of G gradually adds Gaussian noise to transform $\boldsymbol{x}_0^{1:H}$ into $\boldsymbol{x}_T^{1:H} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{I} \in \mathbb{R}^{Hd\times Hd}$ is the identity matrix (Ho, Jain, and Abbeel 2020). The model learns the reverse process by minimizing the reconstruction loss as:

$$\mathcal{L} = \mathbb{E}\left[\|\boldsymbol{x}_0^{1:H} - G(\boldsymbol{x}_t^{1:H}, \boldsymbol{x}_0^{\text{past}}, t, c)\|_2^2\right], \tag{1}$$

where c is a conditioning input that includes a text description and, optionally, the target joint positions. $\boldsymbol{x}_0^{\text{past}}$ represents the preceding motion, and $t \sim \mathcal{U}[1, T]$ denotes the diffusion step. During inference, we apply classifier free guidance (Ho and Salimans 2021), denoted by g, to blend conditional and unconditional predictions:

$$g(\boldsymbol{x}_t^{1:H}) = (1-s)G(\boldsymbol{x}_t^{1:H}, t, \boldsymbol{x}_0^{\text{past}}, \emptyset) + sG(\boldsymbol{x}_t^{1:H}, t, \boldsymbol{x}_0^{\text{past}}, c),$$

where s is a guidance scale. The denoising process progressively refines $\boldsymbol{x}_t^{1:H}$ toward $\boldsymbol{x}_0^{1:H}$.

Physics Based RL Motion Tracking. To track the generated motion plan in simulation, we employ Perpetual Humanoid Control (PHC) (Luo et al. 2023), a general-purpose RL controller for tracking reference motions. The controller π receives as input a state $\boldsymbol{s}^h = (\boldsymbol{s}_{\text{p}}^h, \boldsymbol{s}_{\text{g}}^h)$ at time step h, where $\boldsymbol{s}_{\text{p}}^h$ is the current proprioceptive state, and $\boldsymbol{s}_{\text{g}}^h$ is the goal state. Each motion state includes position $\mathcal{P}$, velocity $\dot{\mathcal{P}}$, rotation $\mathcal{R}$, and angular velocity $\dot{\mathcal{R}}$. The policy $\pi(\boldsymbol{a}^h|\boldsymbol{s}^h) = \mathcal{N}(\mu(\boldsymbol{s}^h), \sigma)$ is modeled as a Gaussian distribution with a fixed diagonal covariance (Luo et al. 2023), where the mean $\mu(\boldsymbol{s}^h)$ is predicted by a neural network. The sampled action $\boldsymbol{a}^h$ is interpreted as a proportional-derivative control target for the SMPL-based humanoid (Loper et al. 2015).

As the motion plan provides only positional information, the goal state is formulated using keypoint imitation: $\boldsymbol{s}_{\text{g}}^h = \boldsymbol{s}_{\text{ref}}^{h+1} \ominus \boldsymbol{s}_{\text{p}}^h := (\mathcal{P}_{\text{ref}}^{h+1} - \mathcal{P}^h, \dot{\mathcal{P}}_{\text{ref}}^{h+1} - \dot{\mathcal{P}}^h, \mathcal{P}_{\text{ref}}^{h+1})$. The policy is optimized using PPO (Schulman et al. 2017) to maximize a composite reward that includes imitation accuracy, energy efficiency (Peng et al. 2018), and stylistic realism via adversarial motion discrimination (Peng et al. 2021).

3.2 Method Overview

We define a long-term task as a sequence of subtasks, with each subtask serving as a minimal behavioral motion unit. Fig. 2 illustrates the overall procedure of BRIC, which consists of three main components: (1) an autoregressive diffusion based motion planner, (2) a RL-based physics controller, and (3) a test-time adaptation mechanism that aligns controller's behavior with the distribution of the motion planner. The adaptation framework in BRIC considers the output distribution of the pretrained RL policy, trained on a large scale human motion dataset (Mahmood et al. 2019), the source domain, and the distribution produced by the diffusion planner as the target domain. To bridge gap between these domains, BRIC updates the policy parameters during inference to better track motion sampled from the planner (Sec. 3.3). In addition to TTA, we introduce a test time guidance strategy that efficiently directs motion plan generation under task specific constraints, enabling rapid and robust adaptation to dynamic environments (Sec. 3.4).

3.3 Test-Time Adaptation of Physics-based Policy to Kinematics Motion

The proposed TTA framework aims to ensure that the policy accurately tracks the generated motion plans. To this end, BRIC addresses two key challenges: (1) preserving the skills of the original policy by mitigating catastrophic forgetting, and (2) reducing error accumulation in the autoregressive policy caused by noise in the motion plan, which may cause physically implausible behavior and tracking failures.

Mitigating Catastrophic Forgetting. At test time, a policy that overfits to the motion plan may suffer from catastrophic forgetting, where it loses previously acquired skills from the source domain (Ebrahimi et al. 2020; Parisi et al. 2019). This problem is particularly critical in our setting, where both the planner and the controller operate in an autoregressive manner. Our policy is trained using PPO (Schulman et al. 2017) and comprises three components: an actor π, a value function V, and a discriminator D used to compute style rewards by distinguishing reference motions from simulated ones (Peng et al. 2021).

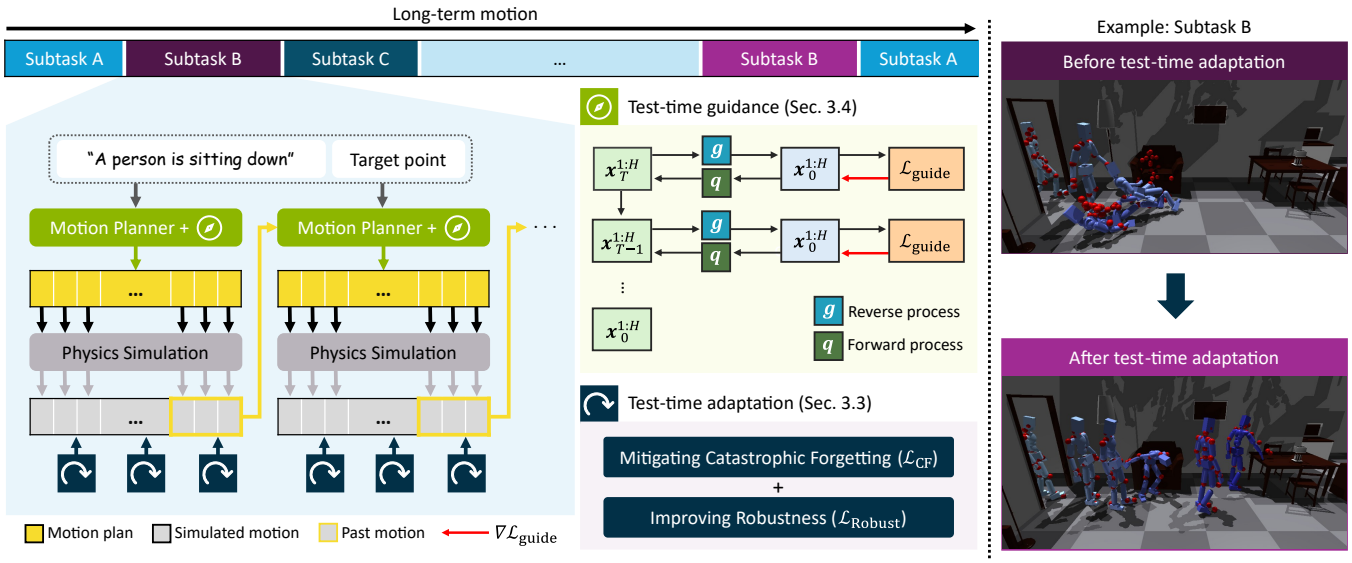

Figure 2: *The overall procedure.* (Left) Test-time adaptation and guidance bridge the distribution gap between the motion planner and physics based controller in an autoregressive manner. (Center) Adaptation mitigates catastrophic forgetting ($\mathcal{L}_{\text{CF}}$) and improves robustness ($\mathcal{L}_{\text{Robust}}$), while guidance refines motion plans by optimizing task objectives. (Right) Before adaptation, the agent fails to track motion plans indicated by red dots. After adaptation, it performs robust and successful motions.

To mitigate catastrophic forgetting during TTA, we introduce a loss that encourages consistency between the current (online) networks and their exponentially averaged counterparts. Following the approach in (Wang et al. 2022), we maintain the online networks, θ_π, θ_V, and θ_D, which are updated through PPO during adaptation. In parallel, we maintain target networks θ'_π, θ'_V, and θ'_D updated via exponential moving average: $\theta'_\pi \leftarrow \alpha\theta'_\pi + (1-\alpha)\theta_\pi$ with a smoothing factor $\alpha \in [0,1]$. These target networks serve as memory buffers that retain prior knowledge for use in TTA. We enforce consistency between the online and target networks using the following loss:

$$\mathcal{L}_{\text{CF}} = \big(V(\boldsymbol{s}^h;\theta_V) - V(\boldsymbol{s}^h;\theta'_V)\big)^2 + \big\|\mu(\boldsymbol{s}^h;\theta_\pi) - \mu(\boldsymbol{s}^h;\theta'_\pi)\big\|_2^2 + \big(D(\boldsymbol{\tau};\theta_D) - D(\boldsymbol{\tau};\theta'_D)\big)^2 \quad (2)$$

where $\boldsymbol{\tau}$ is a ten-frame motion segment, extracted during simulation (Luo et al. 2023; Peng et al. 2021).

Improving Robustness to Noisy Motion Plans. Motion plans generated by diffusion models are not constrained by physical laws and may exhibit various forms of noise, including floating, penetration, or abrupt pose transitions (Yuan et al. 2023; Ren et al. 2023). To enhance robustness against such noise, BRIC introduces a noise-aware loss that measures the KL divergence between the policy distributions conditioned on the executed trajectory and the noisy motion plan. The generated motion plans $\boldsymbol{x}_0^{1:H}$ provide only relative positional information (Guo et al. 2022), whereas the executed trajectory includes both global positions and rotations. To enable comparison, we define a transformation function Ψ that converts the relative coordinates into global one, yielding $\boldsymbol{x}_{\text{plan}}^{1:H} = \Psi(\boldsymbol{x}_0^{1:H})$. We then apply a position-to-rotation network (P2R) (Li et al. 2021) to estimate rotations from the position-only inputs. This process produces global position $\mathcal{P}^h$, rotation $\mathcal{R}^h$, linear velocity $\dot{\mathcal{P}}^h$, and angular velocity $\dot{\mathcal{R}}^h$ for each time step h of the motion plans. Using these quantities, we construct the planned policy input state as $\boldsymbol{s}^h_{\text{plan}} = (\boldsymbol{s}^h_{\text{plan-p}}, \boldsymbol{s}^h_{\text{plan-g}})$, where the current motion state is $\boldsymbol{s}^h_{\text{plan-p}} := (\mathcal{P}^h_{\text{plan}}, \mathcal{R}^h_{\text{plan}}, \dot{\mathcal{P}}^h_{\text{plan}}, \dot{\mathcal{R}}^h_{\text{plan}})$ and the goal state is defined as $\boldsymbol{s}^h_{\text{plan-g}} := (\mathcal{P}^{h+1}_{\text{plan}} - \mathcal{P}^h_{\text{plan}}, \dot{\mathcal{P}}^{h+1}_{\text{plan}} - \dot{\mathcal{P}}^h_{\text{plan}}, \mathcal{P}^{h+1}_{\text{plan}})$. We define the noise in the motion plan as the difference between the planned and executed states: $\boldsymbol{s}^h_{\text{plan}} - \boldsymbol{s}^h$. The noise-aware loss encourages the policy to remain robust under such perturbation, and is defined as the KL divergence:

$$\mathcal{L}_{\text{Robust}} = D_{\text{KL}}(\pi(\boldsymbol{a} \mid \boldsymbol{s}^h) || \pi(\boldsymbol{a} \mid \boldsymbol{s}^h + \beta(\boldsymbol{s}^h_{\text{plan}} - \boldsymbol{s}^h)), \quad (3)$$

where $\beta \in [0,1]$ is a scaling factor that is gradually increased during training to ensure stable learning.

The final loss used to update the motion control policy combines the standard PPO loss with the catastrophic forgetting and robustness objectives as:

$$\mathcal{L}_{\text{PPO-TTA}} = \mathcal{L}_{\text{PPO}} + \lambda_{\text{CF}}\,\mathcal{L}_{\text{CF}} + \lambda_{\text{Robust}}\,\mathcal{L}_{\text{Robust}}, \quad (4)$$

where λ_{CF} and λ_{Robust} are weighting coefficients. Please refer to Appendix B.1 for additional details of the proposed adaptation framework.

3.4 Efficient Test-time Guidance

Existing approaches to test-time guidance aim to optimize a predefined objective function, denoted as $\mathcal{L}_{\text{guide}}(\boldsymbol{x}_{\text{plan}}^{1:H})$, which encodes task-specific constraints (Karunratanakul et al. 2023, 2024; Rempe et al. 2023; Huang et al. 2024). Given a noisy latent sample $\boldsymbol{x}_t^{1:H}$, these methods reconstruct the motion in the signal space as $\boldsymbol{x}_0^{1:H} = g(\boldsymbol{x}_t^{1:H})$, evaluate the objective,

and backpropagate the gradient to the latent space as:

$$\boldsymbol{x}_t^{1:H} \leftarrow \boldsymbol{x}_t^{1:H} - \eta \nabla_{\boldsymbol{x}_t^{1:H}} \mathcal{L}_{\text{guide}}(\Psi(g(\boldsymbol{x}_t^{1:H}))), \quad (5)$$

where η is a step size. The updated latent sample $\boldsymbol{x}_t^{1:H}$ is then denoised to $\boldsymbol{x}_{t-1}^{1:H}$ via the posterior distribution $q(\boldsymbol{x}_{t-1}^{1:H}|\boldsymbol{x}_t^{1:H}, g(\boldsymbol{x}_t^{1:H}))$. This process is repeated iteratively until $t = 1$, gradually injecting the guidance signal into the denoising trajectory.

However, this approach requires costly gradients computation in the latent space, which becomes especially burdensome when the diffusion model G employs a Transformer architecture (Vaswani 2017). To address this inefficiency, we propose a lightweight alternative. Instead of differentiating through the diffusion model, we directly minimize the objective in the signal space by updating the generated motion:

$$\boldsymbol{x}_0^{1:H} \leftarrow \boldsymbol{x}_0^{1:H} - \eta \nabla_{\boldsymbol{x}_0^{1:H}} \mathcal{L}_{\text{guide}}(\Psi(\boldsymbol{x}_0^{1:H})). \quad (6)$$

We then resample the corresponding latent variable using the forward process: $\boldsymbol{x}_t^{1:H} \sim q(\boldsymbol{x}_t^{1:H}|\boldsymbol{x}_0^{1:H})$. Since this method avoids computing gradients with respect to the parameters of the diffusion model G, it offers substantial gains in both memory and computational efficiency. Our guidance method integrates collision loss (Rempe et al. 2023), motion smoothness (Schulman et al. 2014), and heading consistency (Guo et al. 2022) to improve both efficiency and efficacy.

Beyond its standalone advantage, this guidance strategy exhibits strong synergy with the proposed TTA framework. Our signal space optimization in Eq. (6) eliminates the need for backpropagation through G, enabling fast sampling. This efficiency extends to the entire TTA pipeline: when motion plans generated via this lightweight guidance are used as targets for adaptation, the policy update process becomes significantly faster. In effect, fast guidance enables fast adaptation, forming a virtuous cycle. This synergy is a key advantage that improves the practicality of long-term motion generation. The full autoregressive procedure between test-time guidance (TTG) and TTA is provided in Algorithm 1 in Appendix.

4 Experiments

4.1 Setup

Tasks. We evaluate BRIC on four long-term tasks: text-to-motion (T2M), goal-reaching, obstacle avoidance, and indoor human-scene interaction (HSI). For T2M, we follow the FlowMDM protocol (Barquero, Escalera, and Palmero 2024), generating extended motion sequences by composing subtasks corresponding to from 32 text instructions in the HumanML3D test set (Guo et al. 2022).

For goal-reaching, we adopt the setup from CLoSD (Tevet et al. 2025), where the distal target is defined by scaling an initial distance, uniformly sampled from 1 to 3 m, by a random factor in $[1, 100]$, yielding target distances up to 300 m. For obstacle avoidance, we use terrains with obstacles as defined in (Rempe et al. 2023; Rudin et al. 2022), applying the same distance setup as in the goal-reaching task. Both tasks are evaluated across four locomotion styles: *walk*, *walk fast*, *jog*, and *jog slow*.

For the HSI task, we use the largest indoor environment in the ProcTHOR dataset (Deitke et al. 2022), which contains nine interconnected rooms. The agent is required to navigate all rooms and complete four object-interaction subtasks: *REACH*, *SIT*, *GETUP*, and *TOUCH*. We define an HSI scene plan as a sequence of such subtasks. It is generated using ChatGPT-4 (Achiam et al. 2023) from structured combination of text instructions and scene configurations (see Appendix E for further details).

Evaluation Metrics. We evaluate motion expressiveness using R-Precision (R-Prec), Multimodal Distance (MM-Dist), Diversity (Div) and Fréchet Inception Distance (FID) (Guo et al. 2022), and assess transition smoothness using Peak Jerk (PJ) and Area Under the Jerk Curve (AUJ) (Barquero, Escalera, and Palmero 2024). Physical plausibility is measured by counting occurrences of artifacts such as *penetration* (Pen), *floating* (Float), and *foot skating* (Skate) (Yuan et al. 2023). Unlike kinematic models, physics-based simulation may fail to complete motion execution. To capture this, we introduce *execution rate*, defined as the ratio of successfully simulated frames to the total number of frames generated by the motion planner. We further weight PJ and AUJ by this execution rate, which is especially important for evaluating long-term tasks (see Appendix A.2). Finally, we report *success rate* of a task, defined as the fraction of successful trials out of multiple runs. For T2M, results are averaged over ten motion sequences generated per text prompt. For all other task, results are averaged over 1000 simulation trials.

Implementation Details. All experiments are conducted using IsaacGym (Makoviychuk et al. 2021). Following (Tevet et al. 2025), we perform TTA every 32 frames. The condition input c is specified as a text instruction or a combination of a text and a target point. Each motion plan has a length of $H = 60$, with the motion context $\boldsymbol{x}_0^{\text{past}}$ is set to 20 preceding frames. We apply the proposed test-time guidance from Eq. (6) only to the obstacle avoidance and HSI tasks. For obstacle avoidance, only collision detection loss is used (Rempe et al. 2023). For HSI, we generate target point trajectories for *REACH* actions using the A* algorithm (Hart, Nilsson, and Raphael 1968), where collision detection serves as the heuristic function. This trajectory is fed into all evaluated methods. See Appendices A.1 and C.2 for details.

Comparison of Methods. Throughout the experiments, we compare against a baseline model, denoted as *Baseline*, which naively combines PHC and DiP without using TTA. CLoSD is evaluated for all tasks. MaskedMimic (Tessler et al. 2024) is compared only for goal-reaching, as it provides no quantitative results for T2M and states poor performance on long-horizon tasks. UniPhys (Wu et al. 2025a) is compared only on goal-reaching using reported values. UniHSI (Xiao et al. 2024) is compared only on HSI following the evaluation protocol of CLoSD.

4.2 Text-to-Motion

We compare BRIC on the T2M task with state-of-the-art physics-based and kinematic models. Specifically, we evaluate against the physics-based model CLoSD (Tevet et al. 2025), and the kinematic model FlowMDM (Barquero, Escalera, and Palmero 2024). As an additional kinematic reference, we also evaluate our motion planner alone, referred to as *DiP* (Tevet et al. 2025). All diffusion models in this

	Succ. / Exec.	Subsequences				Transition				Physics-based metrics		
		R-prec ↑	FID ↓	Div →	MM-Dist ↓	FID ↓	Div →	PJ →	AUJ ↓	Float ↓	Skate ↓	Pen ↓
GT	-	0.796	0.00	9.34	2.97	0.00	9.54	0.04	0.07	22.9	$206 \cdot 10^{-3}$	0.000
FlowMDM	-	0.685	0.29	9.58	3.61	1.38	8.79	0.06	0.51	21.40	6.64	0.04
DiP	-	0.456	2.30	7.78	5.07	3.27	7.81	0.70	1.80	11.02	7.24	0.07
Baseline	0.113/0.367	0.234	21.63	5.02	7.40	25.23	5.08	0.09 (0.25)	1.07 (5.85)	22.75 (610.23)	2.02 (49.60)	0.03 (0.11)
CLoSD	0.141/0.401	0.243	19.48	5.45	7.29	22.97	5.37	0.09 (0.25)	1.26 (6.21)	22.49 (591.15)	2.81 (47.49)	0.04 (0.13)
BRIC $-\mathcal{L}_{\text{CF}} - \mathcal{L}_{\text{Robust}}$	0.322/0.575	0.277	13.03	6.50	6.87	15.53	6.43	0.07 (0.12)	0.42 (2.08)	22.92 (383.13)	3.13 (36.51)	0.03 (0.09)
BRIC $-\mathcal{L}_{\text{CF}}$	0.428/0.626	0.292	12.06	6.60	6.77	14.16	6.62	0.06 (0.11)	0.30 (1.61)	20.22 (301.69)	4.24 (34.92)	0.02 (0.09)
BRIC $-\mathcal{L}_{\text{Robust}}$	0.441/0.693	0.324	8.54	7.11	6.46	9.55	7.36	0.07 (0.10)	0.36 (1.35)	19.73 (276.44)	2.52 (32.51)	0.04 (0.08)
BRIC (Proposed)	0.494/0.703	0.326	7.96	7.16	6.45	8.90	7.35	0.07 (0.11)	0.45 (1.52)	18.13 (243.61)	3.42 (32.98)	0.04 (0.17)

Table 1: *Comparisons on T2M.* "$-\mathcal{L}_{\text{CF}}$" and "$-\mathcal{L}_{\text{Robust}}$" denote variants of BRIC with the catastrophic forgetting loss in Eq. (2) and robustness loss in Eq. (3) removed, respectively. "$-\mathcal{L}_{\text{CF}} - \mathcal{L}_{\text{Robust}}$" removes both terms and is equivalent to CLoSD finetuned for T2M. "Succ." and "Exec." denote success and execution rates, respectively. Parenthesized numbers show physics and transition metric values weighted by execution rates.

evaluation are conditioned solely on text inputs.

Results. Table 1 shows that BRIC (last row) consistently outperforms CLoSD across all metrics, demonstrating the effectiveness of the proposed TTA in aligning the RL policy with the target motion plan distribution. Notably, BRIC achieves nearly twice the performance of CLoSD in two typically conflicting execution rate and FID (Guo et al. 2022), highlighting its superior robustness and motion quality. As expected, kinematic models perform best in diversity and expressiveness metrics. Nonetheless, BRIC achieves comparable or better scores in PJ and AUJ, which are critical for long-term motions quality (Barquero, Escalera, and Palmero 2024). Moreover, when accounting for failure via execution-weighting metrics, BRIC maintains a significantly larger advantages over CLoSD in both PJ and AUJ. These results indicate that BRIC is well-suited for generating coherent and expressive long-term motion with smooth transitions across diverse motion segments.

4.3 Goal-Reaching and Obstacle Avoidance

We evaluate BRIC against CLoSD, MaskedMimic (Tessler et al. 2024), and UniPhys (Wu et al. 2025a) using success rate as the primary metric. For goal-reaching, we follow the standard UniPhys setup, evaluating performance over two target distance ranges: 1 to 2 m and 3 to 6 m. We also assess a long-term setting as described in Sec. 4.1. For both BRIC and CLoSD, the condition input c includes both text and a target position.

Results. As shown in Fig. 3(a), BRIC consistently achieves the highest success rates, averaged over *walk* and *jog* locomotion styles in the standard setting. Under the long-term setting, we compare BRIC with CLoSD, identified as the strongest existing model, across six distance scales and four locomotion styles: *walk*, *walk fast*, *jog slow*, and *jog*. BRIC demonstrates strong robustness, maintaining a success rate above 0.9 even under the most challenging conditions. In contrast, CLoSD suffers significant performance degradation as distance increases, reaching near-zero success under the largest scale ('scale 100').

For obstacle avoidance, Table 2 shows that both test-time

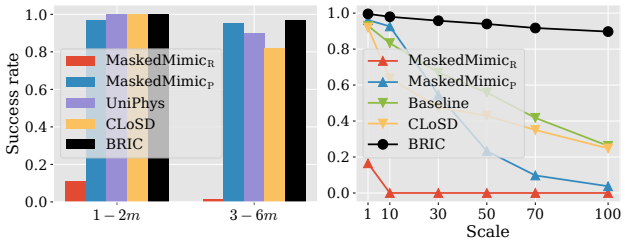

Figure 3: *Comparisons on goal-reaching.* (Left) Standard setting and (Right) long-term setting, each evaluated over two target distance ranges. MaskedMimic$_{\{R,P\}}$ report results for the "reach" and "path-following" modes, respectively, as described in (Tessler et al. 2024).

guidance variants of BRIC perform comparably to one another, and both outperform CLoSD. Implementation details are provided in Appendix C.1.

4.4 Human-Scene Interaction

We evaluate BRIC on a long-term HSI task, comparing it with CLoSD and UniHSI (Xiao et al. 2024). The task follows a predefined scene plan composed of four subtasks, executed within a nine-room indoor environment from the ProcTHOR dataset (Deitke et al. 2022). For the *REACH* subtask, the motion plan is conditioned on either *walk* or *walk fast*, sampled uniformly. The scene plan begins with the humanoid agent entering the environment from the outside (*Enter*), visiting each of the nine rooms sequentially from *Room1* to *Room9*, and exiting the scene by returning to the starting point (*Exit*). In each room, the agent is instructed to navigate toward a target objects while avoiding furniture, and to interact with targets such as sofas or tables with varying geometries. Please refer to Appendix C.2 for details of the HSI task.

Results. Fig. 4 shows that success rate decreases as the agent progresses through rooms for all methods. Rooms such as *Room1*, *Room2*, and *Room9* are particularly challenging due to complex object layouts, requiring precise navigation and robust generalization in object interaction. BRIC main-

Method	Average Success Rate			Computing Cost			
	Scale			TTA		Eval	
	1	10	30	Time	Mem	Speed	Mem
CLoSD	0.80	0.55	0.40	-	-	-	-
Baseline + TTG	0.81	0.60	0.43	-	-	-	-
BRIC + LAT_TTG	**0.96**	**0.83**	0.71	30.24	43574	13.58	34436
BRIC	0.92	**0.83**	**0.76**	**16.21**	**35092**	**27.03**	**12780**

Table 2: *Comparisons on obstacle avoidance.* "Baseline + TTG" applies the proposed test-time guidance to the baseline. "BRIC + LAT_TTG" uses conventional latent space-based test-time guidance (Huang et al. 2024). For each adaptation, elapsed "Time" in seconds, "Speed" in frames per second (FPS), and the GPU memory usage ("Mem") in MiB are reported. Best results are shown in bold.

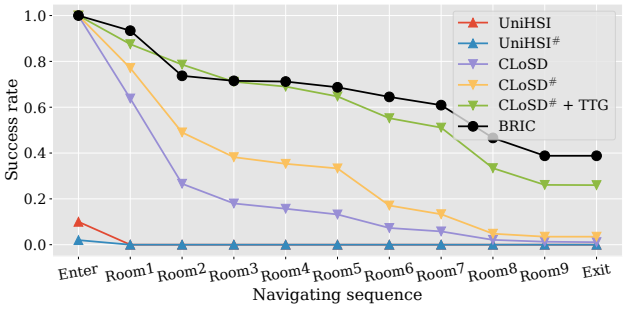

Figure 4: *Comparisons on the HSI task.* The horizontal axis indicates the sequence of rooms visited by the agent according to the scene plan. Models marked with "#" are finetuned for the HSI task (Xiao et al. 2024; Tevet et al. 2025).

tains a success rate of 0.4, outperforming both CLoSD and UniHSI. While CLoSD initially performs better than UniHSI, both degrade to near-zero success, and finetuning in UniHSI does not consistently yield improvements.

Adding test-time guidance with finetuned CLoSD ("CLoSD$^{\#}$+TTG") further improves its performance, but BRIC still achieves the highest final success rate. This indicates that while test-time guidance is effective, motion plans may remain noisy, highlighting the need for adaptive motion control as in BRIC.

4.5 Ablation Study

The bottom section of Table 1 presents an ablation study on the effect of the two regularization terms, $\mathcal{L}_{\text{CF}}$ and $\mathcal{L}_{\text{Robust}}$, used in policy adaptation. Each term individually improves motion expressiveness and physical plausibility, while their combination yields the best overall performance.

Table 2 shows the effect of the proposed test-time guidance. Even without adaptation, the "Baseline+TTG" variant outperforms CLoSD, demonstrating the benefit of lightweight signal-space optimization. Combines TTA to "Baseline+TTG" (BRIC) yields a further increase in success rate, enabled by the adapted policy. Furthermore, our method in Eq. (6) achieves $2.0\times$ faster execution and $2.7\times$ lower memory usage compared to the latent-space gradient method in Eq. (5), enabling significantly more efficient adaptation.

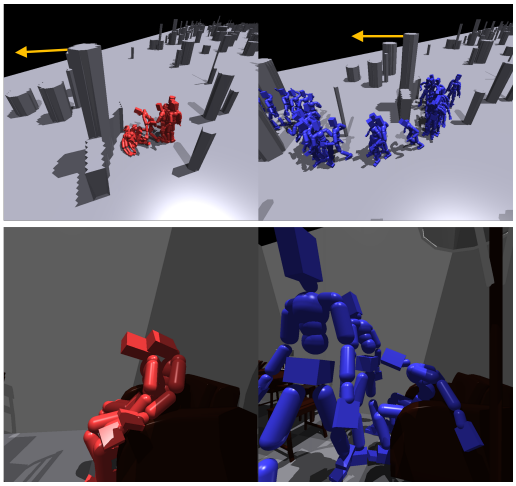

Figure 5: *Qualitative comparison between CLoSD (red) and BRIC(blue).* (Top) In the obstacle avoidance task, CLoSD fails to navigate around obstacles, while our method succeeds under the same conditions. Motion direction is indicated by yellow arrows. (Bottom) In the HSI task, CLoSD becomes stuck and fails to sit, whereas BRICsuccessfully completes *SIT*, *GETUP*, and *REACH* actions (from right to left).

4.6 Qualitative Evaluation

Fig. 1 visualizes full motion trajectories generated by BRIC for long-term obstacle avoidance and HSI tasks. Fig. 5 presents a qualitative comparison between BRIC and CLoSD on these tasks. In obstacle avoidance, CLoSD frequently fails to track the motion plan accurately, leading to collisions or freezing, as shown in Fig. 5(a). Fig. 5(b) highlights a *SIT* action performed on an armchair in *Room2* during the HSI task. CLoSD fails to generalize to the chair's unseen shape, with the agent's arms becoming caught on the armrests. In contrast, BRIC performs interaction successfully, demonstrating the benefit of integrating test-time adaptation and guidance for generalizing to novel object geometries. See Appendix D for additional visualization results.

5 Conclusion

We introduced BRIC, a novel test time adaptation framework for long term motion generation. BRIC enables a physics-based policy to robustly execute noisy motion plans from a diffusion-based kinematic planner via online adaptation. By incorporating a loss that mitigates catastrophic forgetting and a KL divergence auxiliary loss for noise robustness, the policy rapidly adapts while retraining prior skills. Furthermore, we proposed a lightweight test-time guidance method that optimizes task objectives directly in the signal space without backpropagating through the diffusion model, enabling efficient inference control. Extensive experiments shows that BRIC consistently outperforms prior methods in success rates and expressiveness, while preserving physical plausibility.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) through the Korean government (MSIT) under grants (RS-2022-NR069649 and RS-2025-02263810).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Barquero, G.; Escalera, S.; and Palmero, C. 2024. Seamless human motion composition with blended positional encodings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, J.; Hu, P.; Chang, X.; Shi, Z.; Kampffmeyer, M.; and Liang, X. 2025. Sitcom-Crafter: A Plot-Driven Human Motion Generation System in 3D Scenes. In *International Conference on Learning Representations*.

Deitke, M.; VanderBilt, E.; Herrasti, A.; Weihs, L.; Salvador, J.; Ehsani, K.; Han, W.; Kolve, E.; Farhadi, A.; Kembhavi, A.; and Mottaghi, R. 2022. ProcTHOR: Large-Scale Embodied AI Using Procedural Generation. In *Advances in Neural Information Processing Systems*.

Diller, C.; and Dai, A. 2024. Cg-hoi: Contact-guided 3d human-object interaction generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ebrahimi, S.; Meier, F.; Calandra, R.; Darrell, T.; and Rohrbach, M. 2020. Adversarial continual learning. In *European Conference on Computer Vision*.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Hart, P. E.; Nilsson, N. J.; and Raphael, B. 1968. A formal basis for the heuristic determination of minimum cost paths. *IEEE transactions on Systems Science and Cybernetics*, 4(2): 100–107.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*.

Ho, J.; and Salimans, T. 2021. Classifier-Free Diffusion Guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*.

Huang, W.; Jiang, Y.; Wouwe, T. V.; and Liu, K. 2024. Constrained Diffusion with Trust Sampling. In *Advances in Neural Information Processing Systems*.

Janner, M.; Du, Y.; Tenenbaum, J.; and Levine, S. 2022. Planning with Diffusion for Flexible Behavior Synthesis. In *International Conference on Machine Learning*.

Karunratanakul, K.; Preechakul, K.; Aksan, E.; Beeler, T.; Suwajanakorn, S.; and Tang, S. 2024. Optimizing diffusion noise can serve as universal motion priors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Karunratanakul, K.; Preechakul, K.; Suwajanakorn, S.; and Tang, S. 2023. Guided motion diffusion for controllable human motion synthesis. In *IEEE/CVF International Conference on Computer Vision*.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: a skinned multi-person linear model. *ACM Transactions On Graphics*, 34(6): 1–16.

Luo, Z.; Cao, J.; Kitani, K.; Xu, W.; et al. 2023. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *IEEE/CVF International Conference on Computer Vision*.

Makoviychuk, V.; Wawrzyniak, L.; Guo, Y.; Lu, M.; Storey, K.; Macklin, M.; Hoeller, D.; Rudin, N.; Allshire, A.; Handa, A.; and State, G. 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)*.

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

Parisi, G. I.; Kemker, R.; Part, J. L.; Kanan, C.; and Wermter, S. 2019. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113: 54–71.

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

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

Prabhudesai, M.; Ke, T.-W.; Li, A.; Pathak, D.; and Fragkiadaki, K. 2023. Diffusion-tta: Test-time adaptation of discriminative models via generative feedback. *Advances in Neural Information Processing Systems*.

Rempe, D.; Luo, Z.; Bin Peng, X.; Yuan, Y.; Kitani, K.; Kreis, K.; Fidler, S.; and Litany, O. 2023. Trace and pace: Controllable pedestrian animation via guided trajectory diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Ren, J.; Zhang, M.; Yu, C.; Ma, X.; Pan, L.; and Liu, Z. 2023. Insactor: Instruction-driven physics-based characters. *Advances in Neural Information Processing Systems*.

Rudin, N.; Hoeller, D.; Reist, P.; and Hutter, M. 2022. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning*.

Schulman, J.; Duan, Y.; Ho, J.; Lee, A.; Awwal, I.; Bradlow, H.; Pan, J.; Patil, S.; Goldberg, K.; and Abbeel, P. 2014. Motion planning with sequential convex optimization and convex collision checking. *The International Journal of Robotics Research*, 33(9): 1251–1270.

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

Serifi, A.; Grandia, R.; Knoop, E.; Gross, M.; and Bächer, M. 2024. Robot Motion Diffusion Model: Motion Generation for Robotic Characters. In *SIGGRAPH Asia 2024 Conference Papers*.

Song, W.; Zhang, X.; Li, S.; Gao, Y.; Hao, A.; Hou, X.; Chen, C.; Li, N.; and Qin, H. 2024. HOIAnimator: Generating Text-prompt Human-object Animations using Novel Perceptive Diffusion Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Tevet, G.; Raab, S.; Cohan, S.; Reda, D.; Luo, Z.; Peng, X. B.; Bermano, A. H.; and van de Panne, M. 2025. CLoSD: Closing the Loop between Simulation and Diffusion for multi-task character control. In *International Conference on Learning Representations*.

Tevet, G.; Raab, S.; Gordon, B.; Shafir, Y.; Cohen-or, D.; and Bermano, A. H. 2023. Human Motion Diffusion Model. In *International Conference on Learning Representations*.

Tsai, Y.-Y.; Chen, F.-C.; Chen, A. Y.; Yang, J.; Su, C.-C.; Sun, M.; and Kuo, C.-H. 2024. GDA: Generalized diffusion for robust test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wu, Y.; Karunratanakul, K.; Luo, Z.; and Tang, S. 2025a. UniPhys: Unified Planner and Controller with Diffusion for Flexible Physics-Based Character Control. *IEEE/CVF International Conference on Computer Vision*.

Wu, Z.; Li, J.; Xu, P.; and Liu, C. K. 2025b. Human-Object Interaction from Human-Level Instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Xiao, Z.; Wang, T.; Wang, J.; Cao, J.; Zhang, W.; Dai, B.; Lin, D.; and Pang, J. 2024. Unified Human-Scene Interaction via Prompted Chain-of-Contacts. In *International Conference on Learning Representations*.

Xu, M.; Shi, Y.; Yin, K.; and Peng, X. B. 2025a. PARC: Physics-based Augmentation with Reinforcement Learning for Character Controllers. In *SIGGRAPH 2025 Conference Papers*.

Xu, S.; Li, Z.; Wang, Y.-X.; and Gui, L.-Y. 2023. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *IEEE/CVF International Conference on Computer Vision*.

Xu, S.; Ling, H. Y.; Wang, Y.-X.; and Gui, L.-Y. 2025b. InterMimic: Towards Universal Whole-Body Control for Physics-Based Human-Object Interactions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yang, J.; Niu, X.; Jiang, N.; Zhang, R.; and Huang, S. 2024. F-HOI: Toward Fine-grained Semantic-Aligned 3D Human-Object Interactions. In *European Conference on Computer Vision*.

Yi, H.; Thies, J.; Black, M. J.; Peng, X. B.; and Rempe, D. 2024. Generating human interaction motions in scenes with text control. In *European Conference on Computer Vision*.

Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W. T.; and Park, T. 2024. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu, R.; Liu, S.; Yang, X.; and Wang, X. 2023. Distribution shift inversion for out-of-distribution prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yuan, Y.; Song, J.; Iqbal, U.; Vahdat, A.; and Kautz, J. 2023. Physdiff: Physics-guided human motion diffusion model. In *IEEE/CVF International Conference on Computer Vision*.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2024. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6): 4115–4128.

Zhao, K.; Li, G.; and Tang, S. 2025. DART: A Diffusion-Based Autoregressive Motion Model for Real-Time Text-Driven Motion Control. In *International Conference on Learning Representations*.