

Training-free Controllable Human Motion Generation under Heterogeneous Constraints

Xiaofei Hui¹, Yan Bo², Haoxuan Qu¹, Hossein Rahmani¹, and Jun Liu¹

¹ Lancaster University, United Kingdom

² Shenzhen International Graduate School, Tsinghua University, China
{x.hui,h.qu5,h.rahmani,j.liu81}@lancaster.ac.uk, yanb22thu@gmail.com

Abstract. Training-free controllable motion generation has attracted growing interest for enabling flexible constraint enforcement without constraint-specific training. However, existing training-free methods require constraints to be *continuous objective-based* with differentiable losses, while many real-world requirements are *criterion-based* and provide only discontinuous, sparse, or even black-box feedback. In this paper, we propose Motion-Inference-as-Control (MIC), the first training-free motion generation framework that handles both continuous objective-based and criterion-based motion constraints under a shared mechanism. The key idea is to cast diffusion-based motion generation as a stochastic control problem. This perspective not only provides principled and practically effective step-wise control laws that support criterion-based constraints without requiring differentiability and naturally accommodate objective-based constraints as a special case, but also motivates a control-oriented constraint coordination mechanism that adaptively balances and reconciles motion constraints during generation. Experiments across diverse constraint settings demonstrate the effectiveness of our framework.

Keywords: Motion generation · Controllable motion · Text-to-motion diffusion model

1 Introduction

Generating human motion is a fundamental problem in computer vision [23, 61, 76, 84], with broad applications in virtual reality [7, 73], game character animation [35], and humanoid robot control [10]. These applications often require motions to satisfy diverse constraints (e.g., environmental restrictions, user controls, and motion feasibility). A common paradigm to achieve this is to apply constraint-specific model training or adaptation [32, 38, 50, 58, 60, 70, 76]. Yet, as constraint settings can vary substantially across scenes and tasks, such a paradigm may require retraining or adaptation when settings change, which limits scalability.

Recently, training-free controllable motion generation that bypasses constraint-specific model training has attracted growing attention [31, 41]. Specifically, as

 Corresponding Author

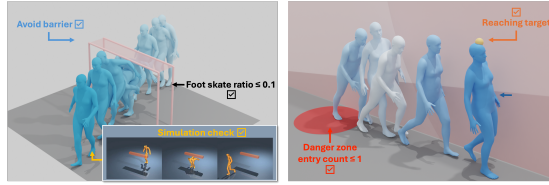


Fig. 1: Real-world motion constraints can be heterogeneous. Some are *continuous objective-based* constraints that can be naturally modeled with differentiable losses (e.g., reaching a target point), whereas many others are *criterion-based* and provide feedback only as sparse, event-triggered, or binary evaluations, such as simulation-based validity tests and safety-threshold rules (e.g., entering the danger zone at most once). In practice, these heterogeneous constraints may be required simultaneously within a single motion. The darker mesh colors indicate later frames in the sequence.

diffusion-based motion generation models [61, 66, 78] offer both high-quality motion generation and substantial inference-time flexibility, recent works [41, 52] have explored enforcing constraints via inference-time modulation on diffusion models, often leveraging gradients computed from differentiable constraints. Yet, despite the promising advances, challenges still remain in effectively integrating diverse real-world constraints in such a training-free manner. Due to the broad range of downstream applications and tasks, motion constraints can be intrinsically heterogeneous and complex, as illustrated in Fig. 1. Some constraints are *continuous objective-based*, naturally expressed as differentiable losses (e.g., reaching target point or maintaining motion velocity), while many constraints are *criterion-based*, providing feedback mainly via event-triggered rules or external evaluators (e.g., safety thresholds, contact events, or feasibility checks) instead of a smooth objective. **This heterogeneity poses the following key challenges.** (1) Criterion-based constraint signals are often discontinuous, sparse, or even black-box (accessible only via query-based evaluation without analytic gradients), making them difficult to incorporate into existing training-free motion generation pipelines that generally rely on differentiable objectives [41, 52]. This calls for an inference-time mechanism capable of handling criterion-based constraints effectively without requiring differentiability. Moreover, as both constraint types can co-exist in practice, this mechanism would ideally also accommodate continuous objective-based constraints under a unified framework, eliminating the need for separate pipelines and facilitating coherent enforcement of heterogeneous constraints. (2) Even with such a mechanism, a further challenge arises when multiple heterogeneous constraints are imposed on the same motion. In practice, these constraints may differ substantially in signal strength and spatial-temporal scope (e.g., full sequence or short time windows, full body or local joints). Naive combination of these heterogeneous signals can lead to *constraint imbalance*, where stronger signals may dominate weaker ones, and *scope interference*, where local constraints may interfere with global ones.

In this paper, we aim to tackle the above challenges and develop a training-free, plug-and-play framework that progressively steers the motion diffusion pro-

cess to satisfy heterogeneous constraints simultaneously and coherently. Essentially, this can be viewed as a step-wise decision problem under motion diffusion dynamics, which decides at each denoising step, how to incorporate constraints into the generation process such that the final generated motion satisfies the constraints. This problem naturally connects to stochastic control [49], which also seeks to determine, at each timestep, how to drive system dynamics toward a desired outcome. Grounded in this connection and inspired by [16, 27, 33, 79], we develop Motion-Inference-as-Control (MIC), a principled and practical framework that formulates inference-time constraint incorporation in diffusion-based motion generation as a stochastic optimal control problem. Notably, this control perspective offers key advantages. Specifically, from stochastic optimal control theory, the desired step-wise diffusion guidance for controllable motion generation can be characterized under a control formulation. For criterion-based constraints, the resulting optimal control law can be expressed using only forward constraint evaluations (e.g., binary feedback and evaluator-dependent feasibility scores), eliminating the dependence on gradient computation. Meanwhile, the same framework naturally encompasses continuous objective-based constraints as a special case, where the optimal control law can also be expressed in the form of gradients. Essentially, the above control-theoretic view provides a unifying principle that integrates heterogeneous constraints into a shared control interface: criterion-based and continuous objective-based constraints alike can be modeled as control signals to steer the motion generation, which directly addresses the first key challenge.

While the above formulation allows heterogeneous motion constraints to be cast as control signals in a unified control interface, naively aggregating multiple such signals can easily lead to constraint imbalance and scope interference, undermining constraint satisfaction and motion coherence. We note that the control-system structure of the above formulation turns this coordination challenge into a more structured design problem. Particularly, it exposes two complementary design axes: how much step-wise control effort to assign to each constraint over time, and how to distribute the resulting effects across joints and frames. Building on this, we propose a bilevel constraint coordination mechanism with two components: a *feedback regulator* that monitors constraint satisfaction and adaptively adjusts the relative importance of each constraint throughout the denoising process; and a *control allocator* that distributes the limited step-wise control budget across constraints according to their effective scopes, seeking to best satisfy all requirements while minimizing inter-constraint interference. Together, this bilevel design mitigates control imbalance and scope interference, enabling stable constraint enforcement while preserving motion naturalness.

Overall, our contributions are: (1) By adopting a control-theoretic perspective, we construct a unified control interface for heterogeneous constraint enforcement in motion generation. To the best of our knowledge, this is the first training-free framework for controllable diffusion-based motion generation that enables both criterion-based and continuous objective-based constraints to be jointly enforced within a single inference-time control formulation. (2) Based on

this foundation, we further propose a practical framework that coordinates different constraint control signals via a constraint coordination mechanism, mitigating control imbalance and scope interference while preserving motion naturalness. (3) Experiment results show that our method consistently achieves strong performance with different constraint settings, demonstrating its effectiveness.

2 Related Work

Text-to-Motion Generation. Text-to-motion generation aims to synthesize realistic human motion sequences conditioned on text descriptions [30, 61, 78, 84]. To tackle this problem, various methods have been proposed, including variational autoencoders (VAEs) [5, 23, 82], generative adversarial networks (GANs) [14, 64], generative masked modeling methods [22, 29], and auto-regressive-based methods [30, 77]. Amid the broad success of diffusion models in generative modeling [18, 20, 39, 40, 71, 72], diffusion-based methods [9, 61, 66, 78] have emerged as a mainstream paradigm for text-to-motion due to their strong ability to generate temporally coherent and diverse motions, motivating a growing body of research building on such diffusion-based models to improve motion generation via both training-based pipelines [2, 19, 25, 37, 60, 76, 83] and training-free mechanisms [8, 11, 47, 53, 75, 85]. In this paper, we focus on leveraging diffusion-based models to improve controllable motion generation in a training-free manner.

Controllable and Constraint-aware Motion Generation. Recent research has explored controllable motion generation that incorporates additional constraint requirements, such as target trajectories, keyframe constraints, and body-part goals [24, 32, 41, 50]. Most existing approaches achieve controllability through task-specific model training or adaptation [6, 28, 36, 38, 50, 59, 65, 70, 81], which, however, can generalize poorly to unseen constraint types [41]. A few efforts have explored leveraging pre-trained motion models to generate constraint-aware motion in training-free manners [31, 41]. These training-free methods depend on gradients computed from differentiable constraint objectives, yet real-world constraints can also be non-differentiable [42, 76]. This gap motivates us to develop a practically effective inference-time framework to integrate such heterogeneous constraints without requiring explicit differentiable formulations.

Control Systems [3, 45] regulate behaviors of dynamical systems via control signals to achieve desired outcomes, and optimal control provides a theoretical framework for designing control laws under system and resource restrictions [62, 74]. As diffusion models essentially define stochastic dynamical processes [26, 27], they can be viewed from a control perspective. Several recent works have explored leveraging control theory for tasks such as music generation [27], image generation [15, 16, 48], video generation [33], and phenotypic response simulation [79]. Different from these methods, we leverage the control-system view to specifically tackle the difficulty of jointly enforcing heterogeneous motion constraints with varying signal properties and spatial-temporal scopes, by developing a practically effective framework that incorporates and reconciles both criterion-based and objective-based constraints during motion generation.

3 Proposed Method: MIC

In this paper, we propose the MIC framework (as shown in Fig. 2), a training-free, plug-and-play constraint control framework that leverages pre-trained motion diffusion models to generate human motions that remain natural while satisfying diverse heterogeneous constraints, in a practically effective manner. To achieve this, diffusion-based controllable motion generation is reinterpreted as a controlled stochastic dynamical process. Based on this, a unified and practically effective step-wise guidance can be developed from optimal control theory: criterion-based constraints are enforced via evaluation feedback without requiring gradients, while continuous objective-based constraints are modeled via gradient-based control (Sec. 3.2). Building on this shared control interface, MIC proposes a constraint coordination mechanism to adaptively balance and reconcile heterogeneous constraints over diffusion steps, reducing constraint imbalance and scope interference when composing multiple constraints (Sec. 3.3).

3.1 Background and Problem Setting

Background on Diffusion-based Motion Generation Process. Given a text prompt, diffusion models [13, 44, 61, 78] generate motion by progressively transforming Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ into a desired motion $\mathbf{x}_0$ via a series of denoising steps in a Markov chain. This is typically modeled as solving reverse-time stochastic differential equations (SDE) [21, 67, 76] from $t = T$ to $t = 0$:

$$d\mathbf{x}_t = \left[-\frac{\beta(t)}{2}\mathbf{x}_t - \beta(t)\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|c^p) \right] dt + \sqrt{\beta(t)}d\bar{\varepsilon}_t, \quad (1)$$

where $\beta(t)$ is the noise schedule of the diffusion process, $\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t|c^p)$ denotes the (estimated) time-dependent score conditioned on a given text prompt c^p , $d\bar{\varepsilon}_t$ is a standard reverse-time Wiener process, and dt represents negative infinitesimal timestep. Note that in Eq. 1, the widely used variance preserving (VP) form of SDE [12, 27, 55] is taken for concreteness, while it can be extended to other SDE variants. Also, the estimated score can be implemented under different denoising parameterizations [26, 32], e.g., by predicting the clean motion state [61] or estimating the noise [78] (see Supplementary for more details).

Problem Setting. In constraint-aware motion generation [31, 32, 41], given a text prompt c^p with a set of target constraints, the model is required to generate a text-conditioned motion sequence $\mathbf{x} \in \mathbb{R}^{N \times J \times D}$ that satisfies the target constraints, where the motion sequence contains N frames of human poses $x_i \in \mathbb{R}^{J \times D}$, each represented by D -dimensional features (e.g., rotations or positions) of J joints. We aim to leverage a pre-trained motion diffusion model as a text-conditioned motion prior $p(\mathbf{x}|c^p)$ and enforce the target constraints in a plug-and-play manner during inference. For constraint target $\mathbf{d}$ and constraint evaluator $v_d(\cdot)$, constraint satisfaction is modeled via an energy-based likelihood $p(\mathbf{d}|\mathbf{x}) \propto \exp(-v_d(\mathbf{x}))$ following [12, 27, 54, 80]. Our goal is to sample motions from the following constraint-aware motion distribution:

$$p(\mathbf{x}|c^p, \mathbf{d}) = \frac{p(\mathbf{x}|c^p)p(\mathbf{d}|\mathbf{x})}{A} \propto p(\mathbf{x}|c^p) \exp(-v_d(\mathbf{x})), \quad (2)$$

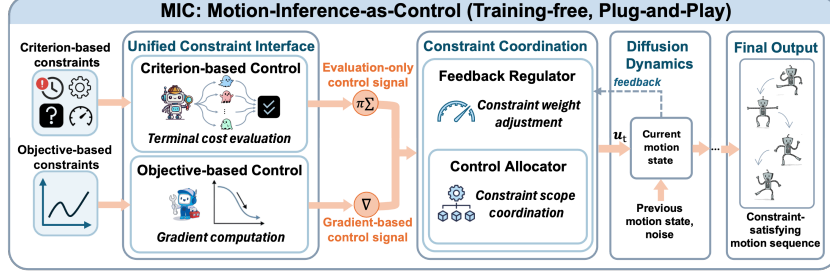


Fig. 2: Illustration of our proposed MIC framework. MIC adopts a unified constraint interface that converts heterogeneous constraints into control signals: criterion-based constraints are handled via forward terminal costs, while continuous objective-based constraints employ gradient computation. A constraint coordination mechanism then integrates these signals, with a feedback regulator adapting constraint weights from violation levels and a control allocator coordinating constraint scopes across joints and frames. The resulting control $\mathbf{u}_t$ is injected at each diffusion step to steer denoising toward constraint satisfaction while minimally perturbing the pretrained motion prior.

where $A = \int_{\mathbf{x}} p(\mathbf{x}|c^p) p(\mathbf{d}|\mathbf{x}) d\mathbf{x}$ is a normalizing constant. Notably, the above distribution defines a target posterior that stays close to the motion prior $p(\mathbf{x}|c^p)$ (thereby maintaining motion naturalness) while biasing samples toward better constraint satisfaction. However, sampling from this posterior can be highly non-trivial in practice: constraint targets $\mathbf{d}$ are often heterogeneous and complex, and the evaluator $v_{\mathbf{d}}(\cdot)$ may provide discontinuous, sparse, or even black-box feedback, making direct constraint enforcement during sampling challenging.

3.2 Optimal Control View of Controllable Motion Generation

Formulating controllable motion generation as optimal control. To tackle the challenges of enforcing heterogeneous constraints in motion generation, inspired by stochastic optimal control [27], we cast inference-time controllable motion generation as an optimal control problem and further develop practically effective instantiations of the resulting control laws for both criterion-based and objective-based constraints. Below, we first show the connection between the motion generation dynamics and the stochastic control process.

To steer the sampling process (Eq. 1) of pre-trained motion diffusion model to instead sample from the desired distribution (Eq. 2), a step-wise guidance term $\mathbf{h}_t(\mathbf{x}_t, t)$ can be injected into the diffusion process, yielding:

$$d\mathbf{x}_t = \left[-\frac{1}{2}\beta(t)\mathbf{x}_t - \beta(t)(\nabla_{\mathbf{x}_t} \log p_t(\mathbf{x}_t | c^p) + \mathbf{h}_t(\mathbf{x}_t, t)) \right] dt + \sqrt{\beta(t)} d\tilde{\epsilon}_t. \quad (3)$$

The problem therefore becomes to find a proper $\mathbf{h}_t$ over time that guides the diffusion dynamics toward the desired posterior $p(\mathbf{x}|c^p, \mathbf{d})$, which is highly non-trivial when the constraints are criterion-based with sparse, non-smooth, or even

black-box feedback. This motivates us to draw on the stochastic optimal control theory to characterize the desired step-wise guidance.

Specifically, Eq. 3 can be interpreted from a control-system view [27, 74]. Intuitively, the diffusion sampling process can be viewed as a stochastic dynamical system evolving from the noisy state $\mathbf{x}_T$ to the final motion $\mathbf{x}_0$, with the step-wise guidance term $\mathbf{h}_t$ acting as control signals that steer the dynamics toward constraint satisfaction. To make this correspondence explicit, define the system state as $\mathbf{z}_t = \mathbf{x}_{T-t}$. The reverse diffusion process can be expressed as the following stochastic dynamical system (more details in Supplementary):

$$d\mathbf{z}_t = \mathbf{b}(\mathbf{z}_t, t)dt + \sigma(\mathbf{z}_t, t)(\mathbf{u}_t(\mathbf{z}_t, t)dt + d\varepsilon_t), \quad (4)$$

where $\mathbf{b}(\mathbf{z}_t, t) = \frac{1}{2}\beta(T-t)\mathbf{z}_t + \beta(T-t)\nabla_{\mathbf{z}_t} \log p_{T-t}(\mathbf{z}_t | c^p)$, $\sigma(t) = \sqrt{\beta(T-t)}$, $\mathbf{u}_t(\mathbf{z}_t, t) = \sqrt{\beta(T-t)}\mathbf{h}_{T-t}(\mathbf{z}_t, T-t)$, $d\varepsilon_t$ is a standard Wiener process, dt is an infinitesimal timestep, and t now denotes forward-time $0 \rightarrow T$. We omit $\mathbf{z}_t$ in $\sigma(\mathbf{z}_t, t)$ and use $\sigma(t)$ in the following, as the noise scale depends only on t .

Then, building on this control process formulation, the problem of finding proper step-wise guidance in diffusion-based controllable motion generation can be cast as an optimal stochastic control problem [27]: finding a per-step *control law* $\mathbf{u}_t^*(\mathbf{z}_t, t)$ that drives the stochastic dynamics (Eq. 4) toward constraint satisfaction, while remaining close to the uncontrolled (prior) motion diffusion dynamics. The optimal control law $\mathbf{u}_t^*$ is obtained by minimizing the *expected cumulative cost* $J_{\mathbf{u}}$ in the control system [74]. In our problem, $J_{\mathbf{u}}$ can be constructed as penalizing motion constraint violations while regularizing deviations from the natural-motion diffusion dynamics:

$$\mathbf{u}_t^*(\mathbf{z}_t, t) = \arg \min_{\mathbf{u}} J_{\mathbf{u}}(\mathbf{z}_t, t) = \arg \min_{\mathbf{u}} \mathbb{E}[\mathcal{E}(\mathbf{z}_T) + \int_t^T L(\mathbf{z}_s, \mathbf{u}_s, s)ds], \quad (5)$$

where $\mathcal{E}(\mathbf{z}_T)$ denotes the terminal cost induced by the motion constraint evaluator $v_{\mathbf{d}}$, penalizing constraint violations measured on the final generated motion $\mathbf{z}_T$, and $\int_t^T L(\mathbf{z}_s, \mathbf{u}_s, s)ds = \int_t^T \frac{1}{2}\|\mathbf{u}_s(\mathbf{z}_s, s)\|^2 ds$ is the running cost that penalizes control effort, encouraging minimal deviation from the motion prior.

Solving the optimal control law. Under the above control formulation, the desired step-wise guidance can be characterized using established results from stochastic optimal control. Specifically, applying the Hamilton–Jacobi–Bellman (HJB) equation [74] and Feynman–Kac formula [46, 74], the optimal control law in Eq. 5 takes the following analytic form [27, 49]:

$$\begin{aligned} \mathbf{u}_t^*(\mathbf{z}_t, t) dt &= \sigma(t) \nabla_{\mathbf{z}_t} \log \psi(\mathbf{z}_t, t) dt \\ &= \sigma(t) \nabla_{\mathbf{z}_t} \log \mathbb{E}_{\mathcal{P}^0}[\exp(-\mathcal{E}(\mathbf{z}_T)) \mid \mathbf{z}_t] dt, \end{aligned} \quad (6)$$

where $\psi(\mathbf{z}_t, t) = \mathbb{E}_{\mathcal{P}^0}[\exp(-\mathcal{E}(\mathbf{z}_T)) \mid \mathbf{z}_t]$ denotes the desirability function associated with the optimal control law $\mathbf{u}_t^*$, and $\mathcal{P}^0$ represents the distribution over diffusion sampling trajectories under the uncontrolled diffusion dynamics, i.e., the text-conditioned natural motion prior dynamics with $\mathbf{u}_t \equiv 0$.

The resulting analytic form of the optimal control law $\mathbf{u}_t^*$ can serve as a principled guidance rule for constraint-aware motion generation. However, this still does not solve the above-mentioned first challenge, as applying the control in Eq. 6 still requires gradient computation, which is unavailable when constraint feedback is provided only via evaluation signals (e.g., discontinuous rules or black-box simulators). This raises a practical question: *can the optimal control law be expressed in a form that does not require gradient computation?*

Solving for criterion-based constraints. Building on stochastic control theory [27, 62], the answer is affirmative. Specifically, the Markov structure of the motion diffusion process and the Gaussian structure of $d\varepsilon_t$ allow the tower rule [68] and Stein’s lemma [57] to express the gradient of the desirability function as a conditional expectation over diffusion noise, yielding an equivalent control law that requires only forward constraint evaluations:

$$\mathbf{u}_t^*(\mathbf{z}_t, t) dt = \frac{\mathbb{E}_{\mathcal{P}^0} [\exp(-\mathcal{E}(\mathbf{z}_T)) d\varepsilon_t | \mathbf{z}_t]}{\mathbb{E}_{\mathcal{P}^0} [\exp(-\mathcal{E}(\mathbf{z}_T)) | \mathbf{z}_t]} \quad (7)$$

Concretely, the resulting solution takes a path-integral form [62] under the uncontrolled diffusion dynamics. As Eq. 7 involves only forward evaluations of the constraint function, it naturally applies to the criterion-based motion constraints, regardless of whether the constraint feedback is intermittent, non-smooth, or only accessible through query-based evaluators. Moreover, *Theorem 1* in the Supplementary shows that applying this control law in motion diffusion sampling recovers the target posterior in Eq. 2.

Having obtained this evaluation-based form for criterion-based constraints, we now turn to the next question: *can we obtain an analytic form of the same control law for continuous objective-based constraints with accessible gradients?*

Solving for continuous objective-based constraints. From Eq. 6, the desirability function $\psi(\mathbf{z}_t, t)$ is defined as the expected constraint score of the final generated motion. When the constraint objective is differentiable, this score is differentiable with respect to the current state, allowing the control law to be expressed in gradient form [27, 49]:

$$\mathbf{u}_t^*(\mathbf{z}_t, t) dt = \sigma(t) \nabla_{\mathbf{z}_t} \log \psi(\mathbf{z}_t, t) dt = \sigma(t) \nabla_{\mathbf{z}} \log p_t(\mathbf{d} | \mathbf{z}_t) dt, \quad (8)$$

where $\mathbf{d}$ is the target constraint conditions and $p_t(\mathbf{d} | \mathbf{z}_t) = \int p(\mathbf{d} | \mathbf{z}_T) p(\mathbf{z}_T | \mathbf{z}_t) d\mathbf{z}_T$ represents the time- t constraint likelihood obtained by marginalizing the terminal likelihood over the uncontrolled diffusion dynamics.

To this end, both criterion-based and continuous objective-based constraints can be represented through corresponding control laws that steer the diffusion sampling dynamics. As Eqs. 7 and 8 are equivalent characterizations of the same underlying optimal control law in Eq. 6, they enter the motion generation dynamics through the same step-wise control interface. This unified control formulation naturally accommodates heterogeneous constraints within a single inference-time framework, enabling principled joint enforcement of diverse constraint types.

Instantiating control signals. Having obtained the optimal control law, we next aim to instantiate it into practical step-wise control signals to apply

to the motion diffusion process. Specifically, as both optimal control forms in Eq. 7 and Eq. 8 involve expectation or marginalization over the uncontrolled diffusion dynamics that are intractable in practice, we need to approximate each form. To derive practically effective instantiation for *criterion-based constraints* where Eq. 7 takes the form of a weighted expectation over $d\varepsilon_t$, we construct the approximation via importance sampling. At each denoising step, we draw M noise increments $\{d\varepsilon_t^{(m)}\}_{m=1}^M$ from a proposal distribution $q = \mathcal{N}(\mu, \Sigma)$, obtain the estimated final motion state $\hat{\mathbf{z}}_T^{(m)}$ via Tweedie’s formula [17], and evaluate the terminal cost $\mathcal{E}(\hat{\mathbf{z}}_T^{(m)})$. We then compute weights according to Eq. 7 and form the step-wise control as:

$$\mathbf{u}_t(\mathbf{z}_t, t)dt = \sum_{m=1}^M \pi_t^{(m)} d\varepsilon_t^{(m)} = \sum_{m=1}^M \frac{\tilde{\pi}_t^{(m)}}{\sum_{j=1}^M \tilde{\pi}_t^{(j)}} d\varepsilon_t^{(m)}, \quad (9)$$

where $\tilde{\pi}_t^{(m)} = \exp(-\mathcal{E}(\hat{\mathbf{z}}_T^{(m)})) \frac{p_0(d\varepsilon_t^{(m)})}{q(d\varepsilon_t^{(m)})}$, with p_0 representing the base diffusion noise distribution. To further improve the approximation, we adapt q to reduce its mismatch with the target distribution $\exp(-\mathcal{E}(\hat{\mathbf{z}}_T))p_0$ via the cross-entropy method [4]. Specifically, at each step, after drawing M samples and evaluating their importance weights $\{\pi_t^{(m)}\}$, we select an elite subset of high-weight samples according to their importance weights and update q using these elite samples [4, 34]. This process can progressively concentrate the proposal on lower-cost regions, thereby increasing the likelihood of sampling noise increments that induce lower terminal costs and improving the practical effectiveness of the approximation. More details are provided in Supplementary.

Meanwhile, for *continuous objective-based constraints*, we instantiate Eq. 8 using a tractable approximation based on Tweedie’s formula [17]. Specifically, at each denoising step, we use Tweedie’s formula to obtain a denoised estimate $\hat{\mathbf{z}}_T$ of the clean motion from the current state $\mathbf{z}_t$, and evaluate the objective-defined gradient on $\hat{\mathbf{z}}_T$ to construct the step-wise control signal:

$$\mathbf{u}_t(\mathbf{z}_t, t)dt = \sigma(t)\nabla_{\mathbf{z}} \log p(\mathbf{d}|\hat{\mathbf{z}}_T)dt. \quad (10)$$

The above approximation aligns with the gradient-based diffusion guidance formula [12]. To this point, we have instantiated practical control signals for both criterion-based and continuous objective-based constraints, which can be injected into the motion generation dynamics through the same mechanism, thereby addressing the first challenge. We next turn to the second challenge: *how to coordinate multiple heterogeneous constraints during the dynamic sampling process*.

3.3 Constraint Coordination Mechanism

In this mechanism, we tackle the challenge where multiple heterogeneous constraints are simultaneously required in generating a certain motion sequence. Specifically, given K motion constraints for a motion sequence, based on Eqs. 9

and 10, we can derive K corresponding control signals $\{\mathbf{u}_{t,k}\}_{k=1}^K$. A straightforward approach to combine them into the final guidance is via a weighted sum $\sum_{k=1}^K W_k \mathbf{u}_{t,k}$, where W_k represents the weight for the k -th constraint. Yet, this alone may not be sufficient for reliable motion generation. In practical tasks, heterogeneous constraints often co-exist, with feedback that varies in signal strength and effective scope over the motion sequence (e.g., a single-frame position target versus a full-sequence constraint). Naively summing these control signals may induce constraint imbalance and interference, degrading motion coherence.

This motivates us to introduce a dedicated coordination mechanism on top of the per-constraint guidance signals. Notably, as we cast all constraint feedback as control signals under a shared control interface, coordinating heterogeneous constraints can be viewed as a structured control-design problem along two complementary axes: how strongly each constraint should act over diffusion steps, and where its effects should be allocated across joints and frames. Inspired by control system design [1, 3], we propose a *constraint coordination mechanism* with two components to stabilize multi-constraint guidance during motion generation: a feedback regulator that adapts constraint strengths, and a control allocator that distributes control effort across joints and frames.

Feedback regulator. To handle constraint imbalance, rather than fixing or manually tuning constraint weights, we treat the weights as closed-loop control variables, as the appropriate weight for a constraint depends on how strongly it is violated, which evolves throughout the denoising process. We therefore draw on the principle of integral feedback control [1]: if a constraint remains violated, its weight should accumulate over time, increasing pressure until the violation is resolved. Specifically, let $\tilde{c}_{k,t} \geq 0$ denote the violation level of the k -th constraint at step t , computed from its evaluator $v_{\mathbf{d}_k}(\cdot)$ on the predicted clean motion at the current step, i.e., $\tilde{c}_{k,t} = \max(0, v_{\mathbf{d}_k}(\hat{\mathbf{z}}_T))$. We normalize $\tilde{c}_{k,t}$ by an exponential moving average-based running scale to obtain $c_{k,t}$, so that violations from heterogeneous constraints are comparable. The feedback regulator then updates $W_{k,t+1}$ (i.e., the weight of the k -th constraint at the $t+1$ -th timestep) as:

$$W_{k,t+1} = \Pi_{[0, W_{\max}]}(W_{k,t} + \gamma c_{k,t}), \quad (11)$$

where γ is the integral gain and $\Pi_{[0, W_{\max}]}$ projects the weight onto a valid range. This integral-style weight adjustment rule has two properties: (1) if a constraint remains violated across steps, its weight increases, applying growing pressure until the violation is resolved; (2) once a constraint becomes satisfied ($\tilde{c}_{k,t} = 0$), its weight stops accumulating and remains stable, preventing it from unnecessarily dominating the control budget. We provide more details in Supplementary.

Control allocator. As each constraint can operate on different scopes of joints and frames, we view the joint-frame blocks in the motion as (spatial-temporal) control channels. This allows constraint composition to be treated as a *budget allocation* problem [3] rather than simply aggregating signals. In particular, each constraint proposes a desired control (i.e., $\mathbf{u}_{k,t}$), but it is only trusted on its effective spatial-temporal scope M_k . The allocator then seeks a single applied control $\mathbf{u}_t$ that best agrees with all constraints on their respective scopes while keeping the total control effort small, thus encouraging minimal deviation

from the pre-trained motion prior. When conflicting constraint control signals are proposed, this allocation step yields a single applied control consensus that reduces cross-scope interference. At step t , the applied control $\mathbf{u}_t$ is computed by solving a weighted least-squares allocation problem:

$$\mathbf{u}_t = \arg \min_{\mathbf{u}} \sum_{k=1}^K \|W_{k,t}(M_k \mathbf{u} - M_k \mathbf{u}_{k,t})\|^2 + \lambda \|\mathbf{u}\|^2, \quad (12)$$

where M_k represents the control scope relevant to the k -th constraint, $W_{k,t}$ encodes its weight at step t (obtained via Eq. 11), and λ is a positive scalar that penalizes the total control effort, preventing excessive deviation from the pre-trained motion prior. Note that for notational convenience, in the above Eq. 12, we flatten the control signal into $\mathbf{u} \in \mathbb{R}^{NJD}$, and accordingly, $M_k \in \mathbb{R}^{NJD \times NJD}$ is a diagonal mask matrix that projects the control signal onto the effective scope of the k -th constraint. This weighted least-squares problem is quadratic and admits a closed-form solution:

$$\mathbf{u}_t = \left(\sum_{k=1}^K W_{k,t}^2 M_k^\top M_k + \lambda I \right)^{-1} \left(\sum_{k=1}^K W_{k,t}^2 M_k^\top M_k \mathbf{u}_{k,t} \right), \quad (13)$$

which can be efficiently computed at each denoising step.

Taken together, by jointly regulating *how strongly* and *where* each constraint influences the motion generation dynamics, MIC enables stable multi-constraint composition while preserving the natural motion prior.

3.4 Overall Inference

Our MIC framework operates in a fully training-free manner. As shown in Fig. 2, given a text prompt c^p and K motion constraints, MIC generates constraint-satisfying and natural motions by iteratively guiding the reverse diffusion sampling process. Specifically, at each denoising step t , MIC (i) computes the control signal $\mathbf{u}_{k,t}$ for each constraint, using Eq. 9 for criterion-based constraints and Eq. 10 for continuous objective-based constraints; (ii) updates the constraint weights via the feedback regulator in Eq. 11; and (iii) resolves interference among constraints via control allocator in Eq. 12 to obtain the applied control $\mathbf{u}_t$, which is injected into the diffusion dynamics to produce the next state. An overall algorithm is provided in Supplementary.

4 Experiments

Benchmark. To evaluate the effectiveness of the proposed MIC framework, we conduct experiments on the very comprehensive training-free controllable motion generation benchmark [41], including two different scenarios: *open-set* (Tab. 1, with control constraints not sourced from motion datasets to evaluate open-set control capability) and *known* (Tab. 2, with control constraints sampled from motion dataset [77]). Following the benchmark, generated motions are

evaluated on a series of requirements (constraints) containing both continuous objective-based and criterion-based ones. The control tasks include human-scene interaction (HSI), geometric constraints (GEO), and human-object interaction (HOI), with HSI-1 for key-frame head-height control, HSI-2 for barrier avoidance, HSI-3 for limited-area walking, GEO-1 for wall touching, and HOI-1 for object moving. Meanwhile, the generated human motions are also required to remain realistic [32, 41], such as maintaining temporal coherence and avoiding foot skating. We also evaluate whether the generated motions can pass a physical simulation check [63]. More details are in the Supplementary.

Metrics. Following [41], we evaluate motion quality using maximum joint acceleration (*Max Acc.*) for frame-wise consistency and foot skating ratio (*Skating*) for motion coherence. We also report the pass rate (*Pass*) of motions that pass physical simulation checks in MuJoCo [63]. For per-task constraint satisfaction, we report constraint error (*C.Err*) and unsucc. rate (*Unsucc. Rate*, percentage of motions failing constraints) following [41]. Additionally, for the *known* scenario (HSI-1 in Tab. 2) where constraint heights are sampled from the HumanML3D test set [77], we evaluate generative quality using FID, Diversity, and R-Precision [52] following [41]. Details are provided in Supplementary.

Implementation Details. We follow the benchmark implementation in [41] to adopt the official Motion Diffusion Model (MDM) pre-trained on HumanML3D dataset [77] and use its DDIM version for fair comparison. Note that our framework can also be applied to other diffusion models. We set the sampling size in Eq. 9 as $M = 16$. The proposal distribution $q = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is initialized as a standard Gaussian and updated at each denoising step using the cross-entropy method [4] with elite ratio 20%. We follow the benchmark implementation to construct the continuous objective-based constraints (e.g., head height constraints) and apply Tweedie’s formula [17] to obtain the denoised estimate $\hat{\mathbf{z}}_T$ at each step. For criterion-based constraints, such as rule-based criteria (e.g., foot-skating and success checks), we evaluate the corresponding constraint criteria to guide the motion generation for our method. For a stable initialization, we adopt the optimization strategy used in prior methods [31, 41] as a warm start, and subsequently apply MIC for step-wise constraint control and coordination. More details about the parameters and implementation are in Supplementary.

4.1 Experiment Results

Quantitative Results. We compare against previous gradient-based methods [31, 41, 52, 67], all using the same MDM backbone. As these methods require differentiable objectives, we use surrogate proxies for criterion-based constraints (e.g., replacing discrete success checks with continuous distance-based scores). As shown in Tab. 1–2, MIC achieves strong overall performance across all tasks. Notably, we observe substantial improvements on criterion-based metrics such as unsucc. rate and simulation check pass rate, while achieving high motion quality. This demonstrates MIC’s effectiveness in handling heterogeneous constraints while maintaining motion generation quality. We provide more details in Supplementary.

Table 1: Comparison with other methods on various open-set constraint specifications following [41]. MDM (Unconstrained) serves as an uncontrollable baseline. All methods are implemented based on the same pre-trained MDM.

Method	Task HSI-2					Task HSI-3				
	Skating↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑	Skating↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑
MDM (Unconstrained) [61]	0.096	0.126	0.454	1.000	0.000	0.096	0.126	0.301	0.875	0.125
IK [41]	0.253	0.923	0.047	0.531	0.500	0.139	0.292	0.015	0.406	0.531
IK+Reg. [41]	0.478	0.352	0.047	0.531	0.469	0.215	0.128	0.015	0.406	0.531
DNO [31]	0.196	0.162	0.051	0.375	0.531	0.144	0.094	0.014	0.344	0.594
ProgMoGen [41]	0.180	0.150	0.097	0.219	0.563	0.125	0.093	0.012	0.344	0.594
ReAlign [67]	0.245	0.155	0.067	0.250	0.469	0.133	0.102	0.016	0.594	0.375
Ours	<i>0.172</i>	<i>0.140</i>	<i>0.009</i>	<i>0.094</i>	<i>0.875</i>	<i>0.112</i>	<i>0.089</i>	<i>0.004</i>	<i>0.250</i>	<i>0.719</i>

Method	Task GEO-1					Task HOI-1				
	Skating↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑	Skating↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑
MDM (Unconstrained) [61]	0.096	0.126	0.233	1.000	0.000	0.029	0.026	1.701	1.000	0.000
MDM Edit [61]	0.161	0.147	0.141	1.000	0.000	0.029	0.032	1.739	1.000	0.000
PriorMDM [52]	0.350	0.197	0.185	1.000	0.000	0.327	0.213	1.884	1.000	0.000
IK [41]	0.147	0.187	0.010	0.219	0.625	0.408	0.919	0.011	0.125	0.875
IK+Reg. [41]	0.536	0.117	0.010	0.219	0.625	0.267	0.153	0.011	0.125	0.875
DNO [31]	0.116	0.113	0.035	0.563	0.375	0.110	0.086	0.035	0.219	0.719
ProgMoGen [41]	0.110	0.104	0.023	0.531	0.406	0.109	0.067	0.028	0.188	0.750
ReAlign [67]	0.122	0.132	0.056	0.625	0.313	0.193	0.169	0.033	0.250	0.625
Ours	<i>0.102</i>	<i>0.088</i>	<i>0.008</i>	<i>0.094</i>	<i>0.781</i>	<i>0.036</i>	<i>0.062</i>	<i>0.004</i>	<i>0.031</i>	<i>0.938</i>

Table 2: Comparison with other methods with constraints sampled from ground-truth HumanML3D test set following [41]. MDM (Unconstrained) serves as an uncontrollable baseline. All methods are implemented based on the same pre-trained MDM.

Method	Task HSI-1					FID↓	Diversity→	R-prec. (Top3)↑
	Skating↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑			
MDM (Unconstrained) [61]	0.086	0.097	0.118	0.718	0.193	0.545	9.656	0.610
MDM Edit [61]	0.094	0.148	0.109	0.645	0.303	0.554	9.656	0.614
IK [41]	0.093	0.414	0.012	0.088	0.735	0.545	9.653	0.610
IK+Reg. [41]	0.269	0.121	0.012	0.088	0.739	0.782	9.509	0.603
DNO [31]	0.085	0.098	0.018	0.121	0.765	0.559	9.595	0.588
ProgMoGen [41]	0.075	0.094	0.012	0.088	0.776	0.556	9.611	0.597
ReAlign [67]	0.096	0.104	0.024	0.189	0.757	0.619	9.433	0.635
Ours	<i>0.074</i>	<i>0.093</i>	<i>0.009</i>	<i>0.068</i>	<i>0.857</i>	<i>0.494</i>	<i>9.656</i>	<i>0.635</i>

Qualitative Results. We also show qualitative comparisons. As shown in Fig. 3, our method produces motions that better satisfy constraints while maintaining motion naturalness and coherence. Moreover, we also incorporate physics-based evaluation [43, 63] to serve as an external constraint evaluator for criteria such as maintaining motion stability. As shown in Fig. 4, while the compared methods often produce unstable motions or fail to satisfy control requirements, our method can generate motions that successfully meet the corresponding requirements.

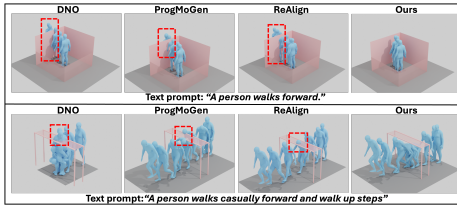


Fig. 3: Visualization comparisons to previous methods. The compared methods produce motions with constraint violations such as boundary penetration and collisions (highlighted in red boxes), whereas our method consistently satisfies constraints while maintaining natural motion quality.

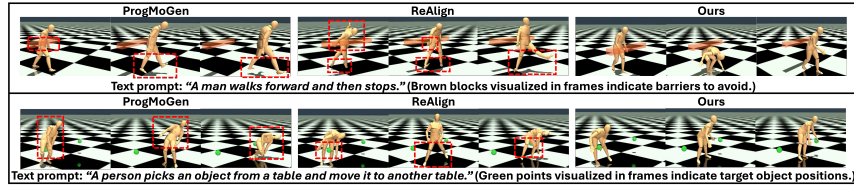


Fig. 4: Qualitative comparisons with ProgMoGen [41] and ReAlign [67] in physical simulation. The motions generated by the compared methods exhibit clear constraint violations and unstable movements (highlighted in red boxes), while MIC consistently satisfies different constraints while maintaining motion naturalness.

4.2 Ablation Study

Below we conduct ablation studies on HSI-2 task. **More ablation studies, further analysis, and user studies are provided in Supplementary.**

Impact of criterion-based constraint formulation. To evaluate our framework for handling criterion-based constraints, we compare against the following four variants. **Baseline A** replaces criterion-based constraints with surrogates and uses gradients to guide diffusion sampling. **Baseline B** estimates gradients of the non-differentiable criteria via zeroth-order optimization [56]. **Baseline C** applies reinforcement learning [69] to approximate gradient guidance for criterion-based constraints. **Baseline D** employs evolutionary strategies [51] to search for effective guidance signals. As shown in Tab. 3, our MIC outperforms all baselines, with a significant improvement on criterion-based constraints such as unsucccess rate. This may be because surrogate proxies can be misaligned with the true criteria, while gradient estimation can be unstable with complex constraints under diffusion dynamics. In contrast, MIC directly constructs control signals for criterion-based constraints without requiring gradients, providing effective guidance.

Impact of unified framework for heterogeneous constraints. Our framework handles criterion-based constraints and continuous objective-based constraints within the same inference time mechanism. To evaluate the benefit of such integration, we compare with the **Separate handling** variant that uses other gradient-based mechanisms in previous methods for continuous objective-based constraints and our mechanism for criterion-based

Table 3: Impact of criterion-based constraint formulation.

Method	Foot Skate↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑
Baseline A	0.180	0.152	0.097	0.219	0.594
Baseline B	0.242	0.267	0.205	0.281	0.594
Baseline C	0.274	0.243	0.231	0.250	0.625
Baseline D	0.208	0.238	0.147	0.219	0.625
MIC	0.172	0.140	0.009	0.094	0.875

Table 4: Impact of unified mechanism.

Method	Foot Skate↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑
Separate handling	0.247	0.165	0.085	0.188	0.656
Unified handling (MIC)	0.172	0.140	0.009	0.094	0.875

constraints and report the best results. As shown in Tab. 4, our unified mechanism achieves better performance, demonstrating its effectiveness.

Impact of the constraint coordination mechanism. In MIC, we propose a constraint coordination mechanism with feedback regulator and control allocator to adaptively coordinate heterogeneous constraints during motion generation. To evaluate this, we conduct experiments with the following variants: **w/o regulation** that only performs control allocation (Eq. 12) without adaptively adjusting the weights of different constraints using feedback regulation; **w/o allocation** that only adjusts the constraint weights via the feedback regulator (Eq. 11) without performing control allocation; and **w/o coordination** that directly averages different control signals without feedback regulation or control allocation. As shown in Tab. 5, MIC achieves the best result, demonstrating the efficacy of our design.

User study. Following [9, 61], we conduct a user study to assess motion quality. We compare MIC with DNO [31], ProgMoGen [41], and ReAlign [67], using 30 motions generated from the same prompts and constraint settings. Twenty participants complete paired comparisons between MIC and each baseline, judging (i) motion naturalness and realism, and (ii) constraint satisfaction, with randomized left-right ordering. Fig. 5 reports the preference rates for MIC, which is consistently favored on both criteria.

Table 5: Impact of the constraint coordination mechanism.

Method	Foot Skate↓	Max Acc.↓	C.Err.↓	Unsucc. Rate↓	Pass↑
w/o regulation	0.208	0.143	0.014	0.125	0.813
w/o allocation	0.189	0.142	0.048	0.125	0.750
w/o coordination	0.220	0.146	0.054	0.156	0.688
<i>MIC</i>	0.172	0.140	0.009	0.094	0.875

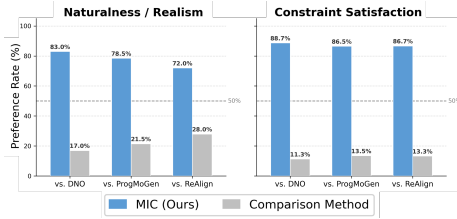


Fig. 5: User study. We report the preference rate (%) of pairwise comparisons between our MIC and each compared method.

5 Conclusion

In this paper, we propose MIC, a training-free framework for guiding pre-trained motion diffusion models to generate natural, constraint-satisfying human motion under heterogeneous real-world constraints. We design a novel inference-time control framework that injects heterogeneous signals into the sampling dynamics via a constraint coordination mechanism that stabilizes and coordinates the constraints. Experiments on tasks with diverse constraint requirements demonstrate the effectiveness of our framework.

References

1. Åström, K.J., Murray, R.: Feedback systems: an introduction for scientists and engineers. Princeton University Press (2021), <https://books.google.co.uk/books?id=qZ0DEAAAQBAJ>
2. Bae, J., Hwang, I., Lee, Y.Y., Guo, Z., Liu, J., Ben-Shabat, Y., Kim, Y.M., Kapadia, M.: Less is more: Improving motion diffusion models with sparse keyframes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11069–11078 (2025)
3. Blaha, T.M., Smeur, E.J.J., Remes, B.D.W.: A survey of optimal control allocation for aerial vehicle control. *Actuators* **12**(7) (2023). <https://doi.org/10.3390/act12070282>, <https://www.mdpi.com/2076-0825/12/7/282>
4. Botev, Z.I., Kroese, D.P., Rubinstein, R.Y., L’ecuyer, P.: The cross-entropy method for optimization. In: Handbook of statistics, vol. 31, pp. 35–59. Elsevier (2013)
5. Cai, Y., Wang, Y., Zhu, Y., Cham, T.J., Cai, J., Yuan, J., Liu, J., Zheng, C., Yan, S., Ding, H., et al.: A unified 3d human motion synthesis model via conditional variational auto-encoder. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11645–11655 (2021)
6. Cao, B., Zheng, S., Wang, Y., Xia, L., Wei, Q., Jin, Q., Liu, J., Lu, Z.: Motionctrl: A real-time controllable vision-language-motion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12253–12262 (2025)
7. Caserman, P., Garcia-Agundez, A., Göbel, S.: A survey of full-body motion reconstruction in immersive virtual reality applications. *IEEE Transactions on Visualization and Computer Graphics* **26**(10), 3089–3108 (2020). <https://doi.org/10.1109/TVCG.2019.2912607>
8. Chen, L.H., Lu, S., Dai, W., Dou, Z., Ju, X., Wang, J., Komura, T., Zhang, L.: Pay attention and move better: Harnessing attention for interactive motion generation and training-free editing. *arXiv preprint arXiv:2410.18977* (2024)
9. Chen, X., Jiang, B., Liu, W., Huang, Z., Fu, B., Chen, T., Yu, G.: Executing your commands via motion diffusion in latent space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 18000–18010 (2023)
10. Cheng, X., Ji, Y., Chen, J., Yang, R., Yang, G., Wang, X.: Expressive Whole-Body Control for Humanoid Robots. In: Proceedings of Robotics: Science and Systems. Delft, Netherlands (July 2024). <https://doi.org/10.15607/RSS.2024.XX.107>
11. Christopher, J.K., Baek, S., Fioretto, F.: Constrained synthesis with projected diffusion models. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) *Advances in Neural Information Processing Systems*. vol. 37, pp. 89307–89333. Curran Associates, Inc. (2024). <https://doi.org/10.52202/079017-2834>, https://proceedings.neurips.cc/paper_files/paper/2024/file/a28af221f2f70be183afc16797a56b91-Paper-Conference.pdf
12. Chung, H., Kim, J., McCann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: The Eleventh International Conference on Learning Representations (2023), <https://openreview.net/forum?id=0nD9zGAGT0k>
13. Dabral, R., Mughal, M.H., Golyanik, V., Theobalt, C.: Mofusion: A framework for denoising-diffusion-based motion synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9760–9770 (2023)
14. Degardin, B., Neves, J., Lopes, V., Brito, J., Yaghoubi, E., Proença, H.: Generative adversarial graph convolutional networks for human action synthesis. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1150–1159 (2022)

15. Domingo-Enrich, C., Drozdal, M., Karrer, B., Chen, R.T.Q.: Adjoint matching: Fine-tuning flow and diffusion generative models with memoryless stochastic optimal control. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=xQBRrtQM8u>
16. Dou, H., Chen, Z., Li, Z., Li, H., Yang, L., Deng, Y.: Constrained particle seeking: Solving diffusion inverse problems with just forward passes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 20870–20878 (2026)
17. Efron, B.: Tweedie’s formula and selection bias. *Journal of the American Statistical Association* **106**(496), 1602–1614 (2011)
18. Foo, L.G., Gong, J., Rahmani, H., Liu, J.: Distribution-aligned diffusion for human mesh recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9221–9232 (2023)
19. Girolamo, M., Lorenzo, M., Bicchierai, M., Berretti, S., Bagdanov, A.D.: No mocap needed: Post-training motion diffusion models with reinforcement learning using only textual prompts. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 967–976 (2026)
20. Gong, J., Foo, L.G., Fan, Z., Ke, Q., Rahmani, H., Liu, J.: Diffpose: Toward more reliable 3d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13041–13051 (2023)
21. Grenander, U., Miller, M.I.: Representations of knowledge in complex systems. *Journal of the Royal Statistical Society: Series B (Methodological)* **56**(4), 549–581 (1994)
22. Guo, C., Mu, Y., Javed, M.G., Wang, S., Cheng, L.: Momask: Generative masked modeling of 3d human motions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1900–1910 (2024)
23. 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 (CVPR). pp. 5152–5161 (June 2022)
24. Guo, Z., Hu, Z., Soh, D.W., Zhao, N.: Motionlab: Unified human motion generation and editing via the motion-condition-motion paradigm. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13869–13879 (2025)
25. Han, G., Liang, M., Tang, J., Cheng, Y., Liu, W., Huang, S.: Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2218–2227. IEEE (2025)
26. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
27. Huang, Y., Ghatare, A., Liu, Y., Hu, Z., Zhang, Q., Sastry, C.S., Gururani, S., Oore, S., Yue, Y.: Symbolic music generation with non-differentiable rule guided diffusion. In: International Conference on Machine Learning. pp. 19772–19797. PMLR (2024)
28. Hwang, I., Bae, J., Lim, D., Kim, Y.M.: Motion synthesis with sparse and flexible keyjoint control. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13203–13213 (October 2025)
29. Jeong, M., Hwang, Y., Lee, J., Jung, S., Kim, W.H.: HGM³: Hierarchical generative masked motion modeling with hard token mining. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=IEul1M5pyk>

30. Jiang, B., Chen, X., Liu, W., Yu, J., Yu, G., Chen, T.: Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems* **36**, 20067–20079 (2023)
31. Karunratanakul, K., Preechakul, K., Aksan, E., Beeler, T., Suwajanakorn, S., Tang, S.: Optimizing diffusion noise can serve as universal motion priors. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1334–1345 (2024)
32. Karunratanakul, K., Preechakul, K., Suwajanakorn, S., Tang, S.: Guided motion diffusion for controllable human motion synthesis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 2151–2162 (October 2023)
33. Kim, J., Kim, B.S., Ye, J.C.: Free2guide: Training-free text-to-video alignment using image lvlm. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 17920–17929 (October 2025)
34. Kroese, D.P., Rubinstein, R.Y., Taimre, T.: Application of the cross-entropy method to clustering and vector quantization. *Journal of Global Optimization* **37**(1), 137–157 (2007)
35. Lee, K., Lee, S., Lee, J.: Interactive character animation by learning multi-objective control. *ACM Trans. Graph.* **37**(6) (Dec 2018). <https://doi.org/10.1145/3272127.3275071>, <https://doi.org/10.1145/3272127.3275071>
36. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: *European Conference on Computer Vision*. pp. 54–72. Springer (2024)
37. Li, Z., Cheng, K., Ghosh, A., Bhattacharya, U., Gui, L., Bera, A.: Simmotionedit: Text-based human motion editing with motion similarity prediction. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 27827–27837 (2025)
38. Li, Z., Luo, M., Hou, R., Zhao, X., Liu, H., Chang, H., Liu, Z., Li, C.: Morph: A motion-free physics optimization framework for human motion generation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 14580–14589 (2025)
39. Li, Z., Rahmani, H., Ke, Q., Liu, J.: Longdiff: Training-free long video generation in one go. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 17789–17798 (June 2025)
40. Li, Z., Rahmani, H., Zhang, J., Xue, Y., Mirmehdi, M., Kuen, J., Gu, J., Liu, J.: Diffgraph: An automated agent-driven model merging framework for in-the-wild text-to-image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 36713–36723 (2026)
41. Liu, H., Zhan, X., Huang, S., Mu, T.J., Shan, Y.: Programmable motion generation for open-set motion control tasks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1399–1408 (2024)
42. Louis, N., Khoshlessan, M., Corso, J.J.: Measuring physical plausibility of 3d human poses using physics simulation. In: *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA (2024), https://bmva-archive.org.uk/bmvc/2024/papers/Paper_615/paper.pdf
43. Luo, Z., Iwase, S., Yuan, Y., Kitani, K.: Embodied scene-aware human pose estimation. *Advances in Neural Information Processing Systems* **35**, 6815–6828 (2022)
44. Meng, Z., Xie, Y., Peng, X., Han, Z., Jiang, H.: Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 27859–27871 (2025)

45. Naidu, D.S.: Optimal control systems. Electrical Engineering Series, CRC press (2018), <https://books.google.co.uk/books?id=Kn50DwAAQBAJ>
46. Oksendal, B.: Stochastic differential equations: an introduction with applications. Universitext, Springer Science & Business Media (2013), <https://books.google.co.uk/books?id=gizqCAAQBAJ>
47. Ota, S., Yu, Q., Fujiwara, K., Ikehata, S., Sato, I.: Pino: Person-interaction noise optimization for long-duration and customizable motion generation of arbitrary-sized groups. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10676–10685 (2025)
48. Pandey, K., Sofian, F.M., Draxler, F., Karaletsos, T., Mandt, S.: Variational control for guidance in diffusion models. In: Forty-second International Conference on Machine Learning (2025), <https://openreview.net/forum?id=Z0ffRRt0im>
49. Pavon, M.: Stochastic control and nonequilibrium thermodynamical systems. Applied Mathematics and Optimization **19**(1), 187–202 (1989)
50. Pinyoanuntapong, E., Saleem, M., Karunratanakul, K., Wang, P., Xue, H., Chen, C., Guo, C., Cao, J., Ren, J., Tulyakov, S.: Maskcontrol: Spatio-temporal control for masked motion synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9955–9965 (2025)
51. Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864 (2017)
52. Shafir, Y., Tevet, G., Kapon, R., Bermano, A.H.: Human motion diffusion as a generative prior. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=dTpbEdN9kr>
53. Shen, Y., Jiang, X., Yang, Y., Wang, Y., Han, D., Li, D.: Understanding and improving training-free loss-based diffusion guidance. Advances in Neural Information Processing Systems **37**, 108974–109002 (2024)
54. Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.Y., Kautz, J., Chen, Y., Vahdat, A.: Loss-guided diffusion models for plug-and-play controllable generation. In: International Conference on Machine Learning. pp. 32483–32498. PMLR (2023)
55. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2021), <https://openreview.net/forum?id=PXTIG12RRHS>
56. Spall, J.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control **37**(3), 332–341 (1992). <https://doi.org/10.1109/9.119632>
57. Stein, C.M.: Estimation of the mean of a multivariate normal distribution. The Annals of Statistics **9**(6), 1135–1151 (1981), <http://www.jstor.org/stable/2240405>
58. Tan, X., Wang, H., Geng, X., Zhou, P.: Sopo: Text-to-motion generation using semi-online preference optimization. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 60714–60747. Curran Associates, Inc. (2025), https://proceedings.neurips.cc/paper_files/paper/2025/file/577cd5863ec73be4e6871340be0936ae-Paper-Conference.pdf
59. Tashakori, A., Tashakori, A., Yang, G., Wang, Z.J., Servati, P.: Flexmotion: Lightweight, physics-aware, and controllable human motion generation. arXiv preprint arXiv:2501.16778 (2025)
60. Tevet, G., Raab, S., Cohan, S., Reda, D., Luo, Z., Peng, X.B., Bermano, A.H., van de Panne, M.: CLoSD: Closing the loop between simulation and diffu-

- sion for multi-task character control. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=pZISppZSTv>
61. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: The Eleventh International Conference on Learning Representations (2023), <https://openreview.net/forum?id=SJ1kSy02jwu>
 62. Theodorou, E., Buchli, J., Schaal, S.: A generalized path integral control approach to reinforcement learning. *The Journal of Machine Learning Research* **11**, 3137–3181 (2010)
 63. Todorov, E., Erez, T., Tassa, Y.: Mujoco: A physics engine for model-based control. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 5026–5033 (2012). <https://doi.org/10.1109/IRCS.2012.6386109>
 64. Wang, Z., Yu, P., Zhao, Y., Zhang, R., Zhou, Y., Yuan, J., Chen, C.: Learning diverse stochastic human-action generators by learning smooth latent transitions. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12281–12288 (2020)
 65. Watanabe, A., Ren, J., Siyao, L., Peng, Y., Wu, E., Simo-Serra, E.: Simdiff: Simulator-constrained diffusion model for physically plausible motion generation. arXiv preprint arXiv:2509.20927 (2025)
 66. Wen, Y., Shuai, Q., Kang, D., Li, J., Wen, C., Qian, Y., Jiao, N., Chen, C., Chen, W., Wang, Y., et al.: Hy-motion 1.0: Scaling flow matching models for text-to-motion generation. arXiv preprint arXiv:2512.23464 (2025)
 67. Weng, W., Tan, X., Wang, J., Xie, G.S., Zhou, P., Wang, H.: Realign: text-to-motion generation via step-aware reward-guided alignment. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 10621–10629 (2026)
 68. Williams, D.: Probability with Martingales. Cambridge mathematical textbooks, Cambridge University Press (1991), <https://books.google.co.uk/books?id=e9saZ0YSi-AC>
 69. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* **8**(3), 229–256 (1992)
 70. Xie, Y., Jampani, V., Zhong, L., Sun, D., Jiang, H.: Omnicontrol: Control any joint at any time for human motion generation. In: The Twelfth International Conference on Learning Representations (2024), <https://openreview.net/forum?id=gd01AEtWso>
 71. Xu, L., Qu, H., Cai, Y., Liu, J.: 6d-diff: A keypoint diffusion framework for 6d object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9676–9686 (2024)
 72. Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.H.: Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.* **56**(4) (Nov 2023). <https://doi.org/10.1145/3626235>, <https://doi.org/10.1145/3626235>
 73. Yin, T., Hoyet, L., Christie, M., Cani, M.P., Pettré, J.: The one-man-crowd: Single user generation of crowd motions using virtual reality. *IEEE Transactions on Visualization and Computer Graphics* **28**(5), 2245–2255 (2022). <https://doi.org/10.1109/TVCG.2022.3150507>
 74. Yong, J., Zhou, X.Y.: Stochastic controls: Hamiltonian systems and HJB equations, vol. 43. Springer Science & Business Media (1999)
 75. Yu, H., Liu, J., Gui, X., Wong, M., Hou, Y., Ong, Y.S.: A plug-and-play multi-criteria guidance for diverse in-betweening human motion generation. arXiv preprint arXiv:2508.01590 (2025)

76. Yuan, Y., Song, J., Iqbal, U., Vahdat, A., Kautz, J.: Physdiff: Physics-guided human motion diffusion model. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 16010–16021 (2023)
77. 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)
78. 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)
79. Zhang, R., Li, C., Lou, Y., Shi, Y., Wang, H., Huang, Y.: Steering where to diffuse: Generative modeling of phenotypic response simulation with steered diffusion bridge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27367–27377 (June 2026)
80. Zhang, S., Zhang, W., Gu, Q.: Energy-weighted flow matching for offline reinforcement learning. In: The Thirteenth International Conference on Learning Representations (2025), <https://openreview.net/forum?id=HA0oLUvuGI>
81. Zhang, Z., Li, Y., Huang, H., Lin, M., Yi, L.: Freemotion: Mocap-free human motion synthesis with multimodal large language models. In: European Conference on Computer Vision. pp. 403–421. Springer (2024)
82. Zhong, C., Hu, L., Zhang, Z., Xia, S.: Att2m: Text-driven human motion generation with multi-perspective attention mechanism. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 509–519 (October 2023)
83. Zhou, W., Dou, Z., Cao, Z., Liao, Z., Wang, J., Wang, W., Liu, Y., Komura, T., Wang, W., Liu, L.: Emdm: Efficient motion diffusion model for fast and high-quality motion generation. In: European Conference on Computer Vision. pp. 18–38. Springer (2024)
84. Zhu, W., Ma, X., Ro, D., Ci, H., Zhang, J., Shi, J., Gao, F., Tian, Q., Wang, Y.: Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(4), 2430–2449 (2024). <https://doi.org/10.1109/TPAMI.2023.3330935>
85. Zhuo, W., Ma, F., Fan, H.: Infinidreamer: Arbitrarily long human motion generation via segment score distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14688–14698 (2025)