ReAlign: Text-to-Motion Generation via Step-Aware Reward-Guided Alignment

Wanjiang Weng[1,2*], Xiaofeng Tan[1,2*], Junbo Wang[3], Guo-Sen Xie[4], Pan Zhou[5], Hongsong Wang[1,2†],

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications (Southeast University), Ministry of Education, China
[3]School of Software, Northwestern Polytechnical University, Xi'an, China
[4]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[5]Singapore Management University, Singapore
{wjweng, xiaofengtan, hongsongwang}@seu.edu.cn, jbwang@nwpu.edu.cn, {gsxiehm, panzhou3}@gmail.com

Abstract

Text-to-motion generation, which synthesizes 3D human motions from text inputs, holds immense potential for applications in gaming, film, and robotics. Recently, diffusion-based methods have been shown to generate more diversity and realistic motion. However, there exists a misalignment between text and motion distributions in diffusion models, which leads to semantically inconsistent or low-quality motions. To address this limitation, we propose **Re**ward-guided sampling **Align**ment (**ReAlign**), comprising a step-aware reward model to assess alignment quality during the denoising sampling and a reward-guided strategy that directs the diffusion process toward an optimally aligned distribution. This reward model integrates step-aware tokens and combines a text-aligned module for semantic consistency and a motion-aligned module for realism, refining noisy motions at each timestep to balance probability density and alignment. Extensive experiments of both motion generation and retrieval tasks demonstrate that our approach significantly improves text-motion alignment and motion quality compared to existing state-of-the-art methods.

Code — https://wengwanjiang.github.io/ReAlign-page

Introduction

With the rising demand for realistic 3D human motions in gaming, filmmaking, virtual reality, and robotics, along with recent advances in motion modeling (Jiang et al. 2023; Wang et al. 2025), there is an increasing need for intuitive and controllable motion generation techniques. Text-to-motion generation, which aims to synthesize human motion directly from natural language descriptions, has emerged as a key research topic (Chen et al. 2023b; Dai et al. 2025; Guo et al. 2022a; Jiang et al. 2023; Wu et al. 2025a,b).

Diffusion has emerged as the mainstream approach for text-driven motion generation (Tevet et al. 2023; Zhang et al. 2024). However, diffusion-based models often struggle with text-motion alignment due to their reliance on text embeddings encoded by CLIP (Radford et al. 2021), which is trained on text-image pairs rather than text-motion pairs. Consequently, these models often fail to capture the semantic alignment between text and motion. The motions synthesized by most existing diffusion-based methods lack coherence with the input descriptions (see Figure 1).

Prior works aiming to improve text-to-motion alignment, such as reinforcement learning with reward functions (Han et al. 2024; Liu et al. 2024; Tan et al. 2025), primarily focus on fine-tuning generative models to enhance motion quality. These approaches lack the capability to handle noisy motion inputs. Moreover, the misalignment issue should be addressed during the denoising process itself, rather than corrected retrospectively after the final motion is generated.

Another limitation of existing diffusion-based methods is that, although they can generate high-quality motions, the generated motions may still lack smoothness and realism in some cases. During the reverse diffusion process, motion generation relies solely on the diffusion model without access to real motion references for guidance. To address this issue, we shift our focus to motion retrieval. In fact, motion generation and retrieval are closely related tasks; however, most existing works investigate them separately, with few efforts dedicated to exploring their interconnection or developing a unified model for both tasks.

To address these problems, we propose a novel **Re**ward-guided sampling **Align**ment strategy (**ReAlign**) to enhance text-motion alignment quality with the guidance of a well-aligned reward distribution. We derive the reward distribution from a step-aware reward comprising two modules: a text-aligned module to ensure semantic consistency, and a motion-aligned module to assess realism. Together, these modules adapt to noisy motions and variations across timesteps, guiding diffusion model toward a distribution that not only maximizes probability density but also maintains strong text-motion alignment. By explicitly addressing both semantic misalignment and motion quality degradation, this approach improves the coherence and realism of the generated motion. The proposed reward model is plug-and-play and can be seamlessly integrated into any motion diffusion model without requiring additional fine-tuning. Our main contributions are as follows:

- **Theoretical reward-guided denoising analysis:** We

*These authors contributed equally.
†Corresponding Author.

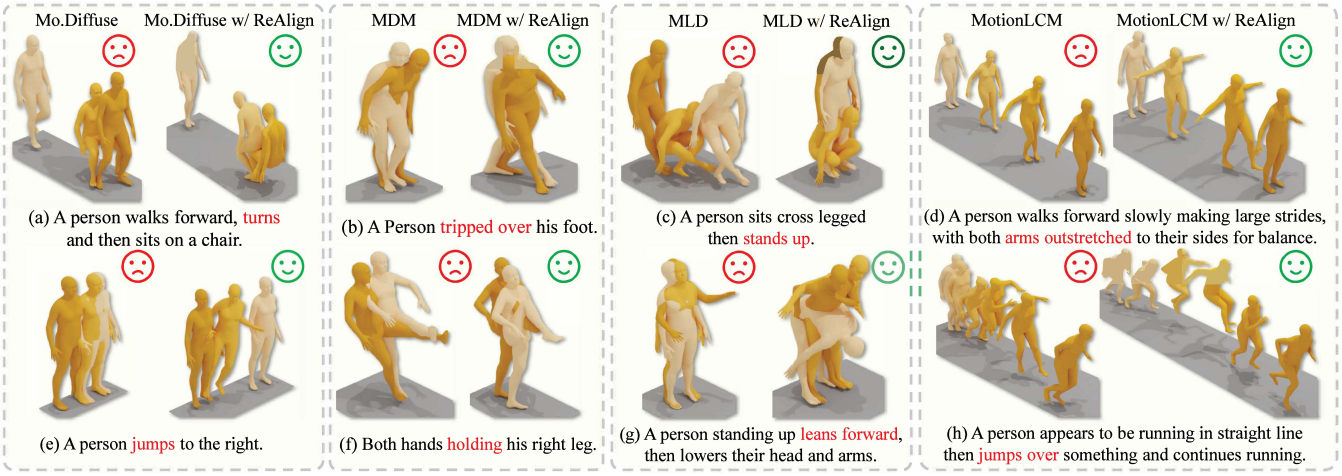

Figure 1: Visual comparison of text-to-motion generation. This figure presents motions generated by existing methods, such as Mo.Diffuse (2024), MDM (2023), MLD (2023b), and MotionLCM (2025). Our ReAlign enhances these models to produce motions that align more closely with text inputs.

theoretically demonstrate that the reward gradient, derived from both text-aligned and motion-aligned rewards, progressively influences the denoising process, guiding the sampling trajectory toward a distribution that better reflects the intended motion semantics.

- **Versatile module for diffusion-based generation:** We propose ReAlign, which comprises a step-aware reward model and a reward-guided sampling strategy to improve text-motion alignment. Extensive experiments demonstrate that our approach significantly enhances existing diffusion-based motion generation models.

Related Works

Alignment in Text-to-Motion Generation. Text-to-motion generation represents a critical task in computer vision, exhibiting rapid advancements in recent years (Zhang et al. 2025a, 2023c,b; Yuan et al. 2025). Diffusion models have been adopted for text-driven motion generation (Tevet et al. 2023; Zhang et al. 2024). MotionLCM (Dai et al. 2025) refines motion-latent diffusion to enable precise spatiotemporal control via few-step inference.

Alignment represents a versatile technique widely employed across the domains of language modeling (Rafailov et al. 2023; Yang et al. 2023), image generation (Wallace et al. 2024), and policy optimization (Chen et al. 2023a). Recently, human preference alignment is studied in text-to-motion generation. ReinDiffuse (Han et al. 2024) refines the diffusion model through reinforcement learning to enhance the physical plausibility of generated motions. MotionRL (Liu et al. 2024) focus aligning human preferences using the proposed multi-reward reinforcement learning framework. SoPo (Tan et al. 2025) combines the strengths of online and offline direct preference optimization to overcome their individual shortcomings, delivering enhanced motion generation quality and preference alignment. However, these methods focus on fine-tuning generative models to align preferences or enhance motion quality without explicitly addressing text-motion misalignment. In contrast, we tackle this issue with a plug-and-play reward model in the inference process.

Diffusion-Based Reward-Guided Generation. Reward guidance for diffusion models can be broadly categorized into derivative-free and gradient-based approaches. Derivative-free methods include SMC-based guidance (Dou and Song 2024) and value-based importance sampling (Li, Huang, and Wei 2025), where the former introduces batch-level interaction while the latter operates independently per sample. Gradient-based guidance is typically instantiated as classifier guidance (Dhariwal and Nichol 2021). For discrete diffusion models, inference-time guidance has been tailored to the discrete setting (Uehara et al. 2025), with practical discrete classifier guidance implementations (Nisonoff et al. 2025) and search-based inference alignment methods (Wan et al. 2024). Diffusion models can also be edited to suppress specific concepts (Wu et al. 2025c). In image synthesis, Liu et al. (2023) further propose a unified multimodal guidance framework with language and image inputs. While reward-guided generation remains relatively under-explored in motion synthesis, this work addresses that gap.

Methods

Motivation and Overview Framework

Preliminaries. Existing diffusion-based motion generation methods (Chen et al. 2023b; Tevet et al. 2023) operate via a forward process and a reverse process. The forward process gradually adds noise into the real motion distribution $p_{\text{data}}(\cdot)$ over timestep, and can be modeled as a stochastic differential equation (SDE) (Song et al. 2021):

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{1}$$

where t is timestep, $\mathbf{f}(\cdot,\cdot)$ and $g(\cdot)$ are the drift and diffusion coefficients, and $\mathbf{w}$ is the Wiener process. For reverse process, motions $\mathbf{x}$ are generated via trajectory sampling (Song

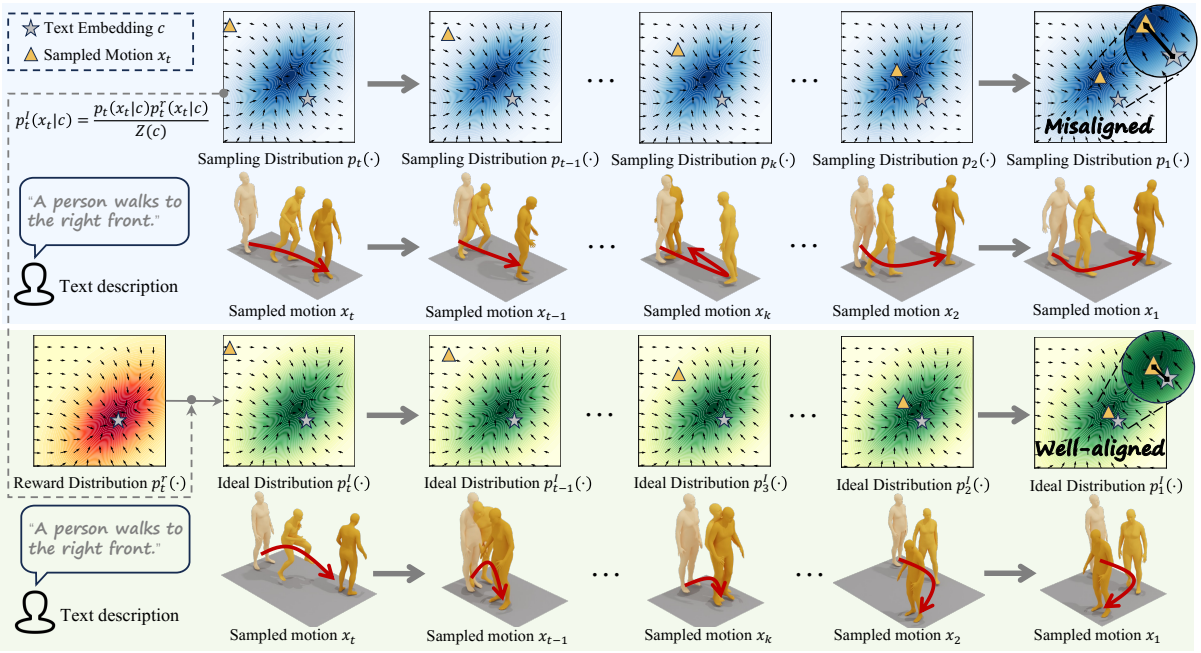

Figure 2: Illustration of the sampling process in diffusion-based motion generation frameworks. The blue region represents the sampling distribution $p_t(\cdot)$ learned by the diffusion model, while the green region depicts the ideal sampling distribution $p_t^I(\cdot)$ achieved by incorporating our proposed reward-guided sampling strategy with the sampling distribution $p_t(\cdot)$.

et al. 2021):

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \quad (2)$$

where $\nabla \log p_t(\mathbf{x})$ is the score function of $p_t(\mathbf{x})$, directing sampling toward higher-density regions.

Motivation. While existing text-to-motion diffusion models enable motion generation with high-quality, they often fail to generate motions that accurately align with textual descriptions. For example, as illustrated in Figure 2, the diffusion model prompted to generate a person walking forward to the right may instead veer left. This misalignment arises as the sampling distribution $p_t(\mathbf{x})$, learned from the diffusion, prioritizes high-probability regions over semantic fidelity.

Upon analyzing the diffusion sampling process (Figure 2), we identify a key issue: sampled motions $\mathbf{x}_t$ (stars) are guided by gradient descent toward high-density regions $p_t(\cdot)$ but consistently diverge from text embeddings c (triangles). This bias prioritizes probability density over semantic alignment, largely due to the reliance on CLIP (Radford et al. 2021) as the text encoder. While aligning text with static images, CLIP struggles with the temporal dynamics of motion, hindering the diffusion model's ability to learn a semantically coherent sampling distribution.

A direct solution is to learn a latent space that aligns motion-text pairs and then train the diffusion model accordingly. However, the scarcity of motion-text pairs makes it difficult to train a generalized text encoder for motion, reducing the diffusion model's generalization ability. Instead, we propose a more effective approach: leveraging an already well-aligned distribution to guide the misaligned sampling process. Accordingly, we first estimate a reward distribution $p_t^r(\mathbf{x})$ from text-motion pairs, capturing semantic alignment. We then integrate this reward distribution with the vanilla sampling distribution to construct an ideal distribution $p_t^I(\mathbf{x})$. Crucially, our method is independent of the diffusion training process, allowing seamless integration into any diffusion model without any finetuning. As shown in Figure 2, sampling from this ideal distribution ensures both high-probability density and strong semantic alignment, overcoming previous limitations.

Overview Framework. Our framework enhances diffusion-based motion generation by constructing an ideal sampling distribution that balances motion probability with text-motion alignment. This section describes how we integrate the reward distribution into the diffusion process and sample from the resulting ideal distribution.

Formally, assume a reward distribution $p_t^r(\mathbf{x}|c)$ has been estimated. Then we define the ideal distribution as:

$$p_t^I(\mathbf{x}|c) = p_t(\mathbf{x}|c)p_t^r(\mathbf{x}|c)/Z(c), \quad (3)$$

where $Z(c) = \int p_t(\mathbf{x}|c)p_t^r(\mathbf{x}|c)\mathrm{d}\mathbf{x}$ is a normalizing constant. This formulation integrates the original sampling distribution $p_t(\mathbf{x}|c)$ with the reward distribution $p_t^r(\mathbf{x}|c)$, balancing both probability density and text-motion alignment.

Using this ideal distribution, we modify the reverse process for trading-off semantic alignment and high-probability sampling as stated in the following theorem.

Theorem 1. *When using the ideal sampling distribution $p_t^I(\mathbf{x}|c)$ in Eq. (3) to replace the vanilla sampling distribution $p_t(\mathbf{x}|c)$, the reverse SDE becomes:*

$$\mathbf{d}\mathbf{x} = \Big[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla \big(\log p_t(\mathbf{x}|c) + \log p_t^r(\mathbf{x}|c)\big)\Big]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \quad (4)$$

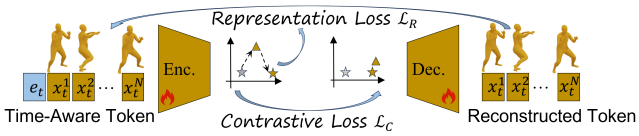

Figure 3: Framework of step-aware reward model. During this process, time-aware tokens, consisting of timestep embedding t and motion embeddings x_t^k, are aligned with text embedding c in the latent space and reconstructed via the decoder, with the encoder and decoder jointly optimized by contrastive loss $\mathcal{L}_C$ and representation loss $\mathcal{L}_R$ (Petrovich, Black, and Varol 2022).

Theorem 1 shows that the gradient of the ideal sampling distribution decomposes into the gradients of $p_t(\mathbf{x}|c)$ and $p_t^r(\mathbf{x}|c)$. Since $p_t(\mathbf{x}|c)$ is already known, the estimated reward distribution can directly guide the sampling process toward the ideal distribution. Next, we detail the estimation of the reward distribution and outline the motion sampling procedure.

Step-Aware Alignment for Reward Distribution

A core challenge in estimating the reward distribution $p_t^r(\mathbf{x}|c)$ is achieving precise motion-text alignment under varying noise levels in the reverse diffusion process (Liang et al. 2025; Rempe et al. 2023). Existing methods (Petrovich, Black, and Varol 2023; Karunratanakul et al. 2023; Li et al. 2025) assume clean and noise-free motion, and overlook timestep-dependent distortions, resulting in coarse and inconsistent alignments. This misalignment hinders accurate reward estimation, which is critical for guiding sampling toward semantically faithful motion generation. To address this, we introduce a step-aware reward model for noise-adaptive alignment and a motion-to-motion reward to ensure consistency with real-world motion patterns implied by text. These components are integrated into a unified reward distribution to enhance alignment and motion quality.

Step-Aware Reward Model. To mitigate timestep-dependent misalignment, we introduce a step-aware reward model $R(\cdot)_\varphi$ illustrated in Figure 3, which explicitly accounts for noise variations across diffusion timesteps. Unlike conventional alignment models (Petrovich, Black, and Varol 2023; Li et al. 2025), our approach incorporates a timestep token $[e_t]$ into the motion representation, allowing the model to learn noise-dependent alignment patterns. Given an N-frame motion sequence $[x_t^1, x_t^2, \ldots, x_t^N]$, we augment it with the timestep token to form the enriched representation $[e_t, x_t^1, x_t^2, \ldots, x_t^N]$. This enables the transformer-based encoder to process motion dynamics while adapting to different noise levels.

During training, noise is added to motion at timestep t, and the step-aware reward model $R_\varphi(\mathbf{x}_t, c)$ is optimized by two complementary losses: a representation loss $\mathcal{L}_R$ (Petrovich, Black, and Varol 2022) to learn meaningful motion embeddings, and a contrastive loss $\mathcal{L}_C$ (Oord, Li, and Vinyals 2018) to ensure accurate motion-text retrieval. The overall training loss $\mathcal{L}_{RM}(\varphi; \mathbf{x}_t, c)$ is defined as:

$$\mathcal{L}_{RM}(\varphi; \mathbf{x}_t, c) = \mathcal{L}_C(\varphi; \mathbf{x}_t, c) + \mathcal{L}_R(\varphi; \mathbf{x}_t, c). \quad (5)$$

Algorithm 1 detail the training procedure of the step-aware reward model.

Algorithm 1: Training Step-Aware Reward Model

Input: Step-aware reward model R_φ, training set $\mathcal{D}_{tr}$, timestep T range $[t_{\min}, t_{\max}]$, probability parameter ω, noise scheduler α.
Output: Step-aware reward model R_φ.

repeat
for $(\mathbf{x}, c)$ in $\mathcal{D}_{tr}$ **do**
  $t \leftarrow 0$ ▷ Initialize t
  if Uniform(0,1) $> \omega$ **then**
    $t \leftarrow$ Uniform$(t_{\min}, t_{\max})$ ▷ Add noise to motion
  end if
  $\mathbf{x}_t \sim \mathcal{N}(\sqrt{\alpha}_t \mathbf{x}, (1-\alpha_t)\mathbf{I})$ ▷ Forward process
  $\mathcal{L}_{RM}(\varphi; \mathbf{x}_t, c)$ by Eq. (5) ▷ Compute loss
  $\varphi \leftarrow \varphi - \nabla_\varphi \mathcal{L}_{\text{RM}}(\varphi)$ ▷ Update parameter
end for
until converged

Once trained, the step-aware reward model establishes a well-aligned latent space. Given a motion $\mathbf{x}$ and text condition c, it evaluates their semantic alignment as:

$$R_\varphi(\mathbf{x}, c) = \cos(\mathbf{z}_\mathbf{x}, \mathbf{z}_c), \quad (6)$$

where $\mathbf{z}_\mathbf{x}$ and $\mathbf{z}_c$ are the respective motion and text embeddings in the learned latent space.

Motion-to-Motion Reward. While text-to-motion alignment is essential, text descriptions often exhibit ambiguity, leading to inconsistencies in generated motions. To mitigate this, we introduce a motion-to-motion reward, which evaluates alignment by comparing the generated motion $\mathbf{x}_t$ with a reference motion $\mathbf{x}^c$ retrieved from the training set $\mathcal{D}_{tr}$. The step-aware reward model is used to select $\mathbf{x}^c$ as the closest match to the text condition c:

$$\mathbf{x}^c = \arg\max_{\mathbf{x} \in \mathcal{D}_{tr}} R_\varphi(\mathbf{x}, c). \quad (7)$$

This retrieved motion $\mathbf{x}^c$ acts as a dynamic anchor, ensuring that generated motions remain faithful to real-world motion patterns implied by the text. Accordingly, The motion-aligned reward is then computed as:

$$R_m(\mathbf{x}_t, c) = \cos(\mathbf{z}_\mathbf{x}, \mathbf{z}_{\mathbf{x}^c}), \quad (8)$$

where $\mathbf{z}_\mathbf{x}$ and $\mathbf{z}_{\mathbf{x}^c}$ are the embeddings of the generated and retrieved motions, respectively. This ensures generated motions adhere to real-world motion patterns while maintaining semantic consistency.

Reward Distribution. With both the step-aware reward model and the motion-to-motion reward, we define the dual-alignment reward as:

$$R(\mathbf{x}_t, c) = \mu R_\varphi(\mathbf{x}_t, c) + \eta R_m(\mathbf{x}_t, c), \quad (9)$$

where μ and η control the contributions of text-based and motion-based alignment. This reward formulation defines the reward distribution over noised motion as:

$$p_t^r(\mathbf{x}_t|c) = \exp(R(\mathbf{x}_t, c)) / Z^r(c). \quad (10)$$

Algorithm 2: Reward-Guided Denoise Process

Input: Diffusion model ϵ_θ, reward model R, training set $\mathcal{D}_{tr}$, condition c, timestep T.
Output: Generated motion $\mathbf{x}_0$.

$\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
$\mathbf{x}^c = \arg\max_{x \in \mathcal{D}_{tr}} R_\varphi(\mathbf{x}, c)$
for $t = T, \cdots, 1$ **do**
  use $\mathbf{x}^c$ to obtain reward score
  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ **if** $t > 1$ **else** $\epsilon = \mathbf{0}$
  use Eq. (13) to generate $\mathbf{x}_{t-1}$
end for
return $\mathbf{x}_0$

Here, $Z^r(c) = \int \exp(R_\varphi(\mathbf{x}, c))\mathrm{d}\mathbf{x}$ is for normalization.

By integrating text-motion and motion-motion alignment, our reward signal ensures semantic consistency and coherence, guiding diffusion sampling to produce high-fidelity, text-faithful motions.

Reward-Guided Sampling

Building on the dual-alignment reward $R(\mathbf{x}_t, c)$ and its associated distribution $p_t^r(\mathbf{x}_t|c)$, we now integrate them into the reverse SDE to refine motion generation. The following theorem establishes how this reward distribution enhances sampling for precise text-conditioned synthesis.

Theorem 2. *Given the reward distribution $p_t^r(\mathbf{x}|c)$ defined in Eq. (10), the reverse SDE can be rewritten as:*

$$\mathbf{d}\mathbf{x} = \Big[\mathbf{f}(\mathbf{x}, t) - g(t)^2 \nabla \big(\log p_t(\mathbf{x}|c) + R\left(\mathbf{x}_t, c\right)\big)\Big]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{11}$$

Theorem 2 reveals that the reward gradient $\nabla R(\mathbf{x}_t, c)$, derived from both text-aligned and motion-aligned reward components, directly influences the sampling trajectory. Integrating these gradients into the reverse SDE can dynamically steer the sampling toward a distribution that better aligns with both textual conditions and realistic structures.

Building upon this continuous-time formulation, for practical motion generation we then derive its discrete approximation within the DDPM (Ho, Jain, and Abbeel 2020) framework in the following theorem.

Theorem 3. *Given a reverse SDE defined in Eq. (11), adopting standard DDPM settings (Song et al. 2021; Ho, Jain, and Abbeel 2020) where $\mathbf{f}(\mathbf{x}, t) = -\frac{1}{2}\bar{\beta}_{t+\Delta t}\mathbf{x}_t$, $g(t) = \sqrt{\beta_{t+\Delta t}}$, and $\bar{\beta}_t = \frac{\beta_{t+\Delta t}}{\Delta t}$, with time steps $N \to \infty$ and step size $\Delta t = \frac{1}{N}$, the reward-guided denoising process is given by:*

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\bar{\mathbf{x}}_{t-1} + \sqrt{\beta_t}\epsilon\right) + \frac{\beta_t}{\sqrt{\alpha_t}}\nabla R\left(\mathbf{x}_t,\, c\right), \tag{12}$$

where $\bar{\mathbf{x}}_{t-1} = \mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t, c)$, β_t and α_t are the noise schedule parameters, $\epsilon_\theta(\cdot)$ represents the diffusion model, and ϵ is Gaussian noise sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

Theorem 3 demonstrates that the reward gradient $\nabla R(\mathbf{x}_t, c)$, weighted by $\frac{\beta_t}{\sqrt{\alpha_t}}$, progressively influences the denoising process, adapting the sampling trajectory toward a distribution that reflects the intended motion semantics.

To ensure the sampling stability, we remove the weight $\frac{\beta_t}{\sqrt{\alpha_t}}$ on the reward term, leading to a revised denoising process:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\bar{\mathbf{x}}_{t-1} + \sqrt{\beta_t}\epsilon\right) + \nabla R(\mathbf{x}_t, c). \tag{13}$$

Based on the theoretical framework above, we propose Algorithm 2, which integrates the step-aware reward model with off-the-shelf classifier-free guidance (CFG) into the diffusion-based generation process(Ho and Salimans 2022).

Experiment

Experiment Setting

Datasets and Evaluation Metrics. We employ two widely used text-to-motion datasets, HumanML3D (Guo et al. 2022a) and KIT-ML (Plappert, Mandery, and Asfour 2016) for evaluation purposes. Consistent with the majority of prior studies (Guo et al. 2024; Li et al. 2025), we adopt R-Precision for Top k, Fréchet Inception Distance (FID), Multi-Modal Distance (MM Dist), and Diversity as evaluation metrics to assess the generation quality and alignment accuracy of our model.

Implementation Details. We employ the SkipTransformer (Chen et al. 2023b) as the foundational architecture for our step-aware reward model, consisting of a transformer encoder processing both text and motion inputs, alongside a motion decoder. Each component features 9 layers and 4 attention heads, with the latent space dimension fixed at 256. The training process incorporates a maximum timestep of 1000, a noisy motion probability of 0.5, and a negative filtering threshold of 0.9 to regulate the selection of negative samples. For model training, we adhere to the TMR framework (Petrovich, Black, and Varol 2023), employing a composite loss function expressed as a weighted combination $\mathcal{L}_C + \mathcal{L}_R$. Optimization is performed using the AdamW (2017), configured with a learning rate of 10^{-4} and a batch size of 512, while other hyperparameters are consistent with those specified in the TMR (2023).

Results of Motion Generation and Retrieval

Text-to-Motion Generation. As shown in Table 1, our reward-guided sampling, i.e., ReAlign, significantly enhances performance when integrated with state-of-the-art text-to-motion models. Specifically, by integrating our ReAlign, MLD++ (Dai et al. 2025) achieves new SoTA results, with an R@3 of 85.2% (+2.8%), alongside a reduction in FID of 0.055 (+24.7%) and an MM Dist to 2.648 (+5.8%). Furthermore, our ReAlign also significantly enhances the performance of MDM (Chen et al. 2023b), yielding SoTA results on the KIT-ML dataset, with an R@3 of 78.4% (+7.3%), alongside a reduction in FID of 0.276 (+44.5%) and an MM Dist to 2.775 (+10.4%). These consistent improvements over the baseline without ReAlign demonstrate the effectiveness of our reward-guided sampling in enhancing text-motion alignment quality.

Motion-Text Retrieval. Following the small-batch protocol of TMR(Petrovich, Black, and Varol 2023), we evaluate our ReAlign on text–motion retrieval and benchmark it against

Method	R Precision ↑			FID ↓	MM Dist ↓	Diversity →
	Top 1	Top 2	Top 3			
Real	0.511	0.703	0.797	0.002	2.974	9.503
T2M (2022a)	$0.455^{\pm0.002}$	$0.636^{\pm0.003}$	$0.736^{\pm0.003}$	$1.087^{\pm0.002}$	$3.347^{\pm0.008}$	$9.175^{\pm0.002}$
MDM (2023)	$0.455^{\pm0.006}$	$0.645^{\pm0.007}$	$0.749^{\pm0.006}$	$0.489^{\pm0.047}$	$3.330^{\pm0.25}$	$9.920^{\pm0.083}$
T2M-GPT (2023a)	$0.492^{\pm0.003}$	$0.679^{\pm0.002}$	$0.775^{\pm0.002}$	$0.141^{\pm0.005}$	$3.121^{\pm0.009}$	$9.722^{\pm0.082}$
ReMoDiffuse (2023b)	$0.510^{\pm0.005}$	$0.698^{\pm0.006}$	$0.795^{\pm0.004}$	$0.103^{\pm0.004}$	$2.974^{\pm0.016}$	$9.018^{\pm0.75}$
Mo.Diffuse (2024)	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.775^{\pm0.001}$	$0.630^{\pm0.001}$	$3.113^{\pm0.001}$	$9.410^{\pm0.049}$
OMG (2024)	-	-	$0.784^{\pm0.002}$	$0.381^{\pm0.008}$	-	$9.657^{\pm0.085}$
MotionLCM (2025)	$0.502^{\pm0.003}$	$0.698^{\pm0.002}$	$0.798^{\pm0.002}$	$0.304^{\pm0.012}$	$3.012^{\pm0.007}$	$9.607^{\pm0.066}$
Mo.Mamba (2025b)	$0.502^{\pm0.003}$	$0.693^{\pm0.002}$	$0.792^{\pm0.002}$	$0.281^{\pm0.011}$	$3.060^{\pm0.000}$	$9.871^{\pm0.084}$
CoMo (2024)	$0.502^{\pm0.002}$	$0.692^{\pm0.007}$	$0.790^{\pm0.002}$	$0.262^{\pm0.004}$	$3.032^{\pm0.015}$	$9.936^{\pm0.066}$
ParCo (2025)	$0.515^{\pm0.003}$	$0.706^{\pm0.003}$	$0.801^{\pm0.002}$	$0.109^{\pm0.005}$	$2.927^{\pm0.008}$	$9.576^{\pm0.088}$
MARDM (2024)	$0.500^{\pm0.004}$	$0.695^{\pm0.003}$	$0.795^{\pm0.003}$	$0.114^{\pm0.007}$	-	-
MG-MotionLLM (2025b)	$0.516^{\pm0.002}$	$0.706^{\pm0.002}$	$0.802^{\pm0.003}$	$0.303^{\pm0.010}$	$2.952^{\pm0.009}$	$9.960^{\pm0.073}$
EnergyMoGen (2025)	$0.526^{\pm0.003}$	$0.718^{\pm0.003}$	$0.815^{\pm0.002}$	$0.176^{\pm0.006}$	$2.931^{\pm0.007}$	$\mathbf{9.500}^{\pm0.091}$
MLD (2023b)	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$0.473^{\pm0.013}$	$3.196^{\pm0.010}$	$9.724^{\pm0.082}$
w/ ReAlign (Ours)	$0.567^{\pm0.003}$ **(+17.9%)**	$0.759^{\pm0.003}$ **(+12.8%)**	$0.848^{\pm0.003}$ **(+9.8%)**	$0.195^{\pm0.005}$ **(+58.8%)**	$2.704^{\pm0.007}$ **(+15.4%)**	$9.474^{\pm0.068}$ **(+86.9%)**
MLD++(2025)	$0.548^{\pm0.003}$	$0.738^{\pm0.003}$	$0.829^{\pm0.002}$	$0.073^{\pm0.003}$	$2.810^{\pm0.008}$	$9.658^{\pm0.089}$
w/ ReAlign (Ours)	$\mathbf{0.572}^{\pm0.002}$ **(+4.4%)**	$\mathbf{0.764}^{\pm0.002}$ **(+3.5%)**	$\mathbf{0.852}^{\pm0.001}$ **(+2.8%)**	$\mathbf{0.055}^{\pm0.003}$ **(+24.7%)**	$\mathbf{2.648}^{\pm0.008}$ **(+5.8%)**	$9.478^{\pm0.055}$ **(+83.9%)**

Table 1: Comparison of text-to-motion generation performance on the HumanML3D dataset. These metrics are evaluated by the evaluator from TM2T (Guo et al. 2022b). The arrows ↑, ↓, and → indicate higher, lower, and closer-to-real-motion values are better, respectively. **Bold** highlights the best results. Percentages in brackets indicate improvements over respective baselines.

Method	R Precision ↑			FID ↓	MM Dist ↓	Div. →
	Top 1	Top 2	Top 3			
Real	0.424	0.649	0.779	0.031	2.788	11.08
T2M (2022a)	0.361	0.559	0.681	3.022	2.052	10.72
MLD (2023b)	0.390	0.609	0.734	0.404	3.204	10.80
T2M-GPT (2023a)	0.416	0.627	0.745	0.514	3.007	10.86
CoMo (2024)	0.422	0.638	0.765	0.332	2.873	10.95
Mo.Mamba (2025b)	0.419	0.645	0.765	0.307	3.021	11.02
ParCo (2025)	0.430	0.649	0.772	0.453	2.820	10.95
Mo.Diffuse (2024)	0.417	0.621	0.739	1.954	2.958	**11.10**
w/ ReAlign (Ours)	0.419	0.639	0.764	0.805	2.801	10.66
MDM (2023)	0.403	0.606	0.731	0.497	3.096	10.74
w/ ReAlign (Ours)	**0.451**	**0.664**	**0.784**	**0.276**	**2.775**	10.76

Table 2: Comparison of text-to-motion generation performance on the KIT-ML dataset. **Bold** highlights the best results, Div. stands for Diversity.

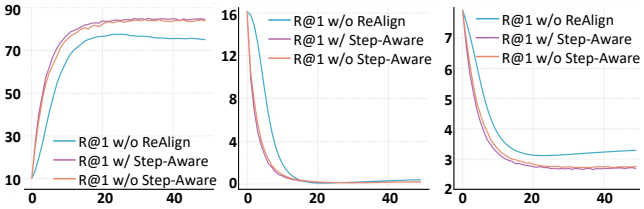

Figure 4: Comparison of motion generation quality across denoising steps for the MLD w/o ReAlign, MLD w/o Step-Aware, and MLD w/ Step-Aware (ReAlign).

the latest state of the art. As summarized in Table 3, ours attains 67.59% R@1 and 87.44% R@3 on HumanML3D for the text-to-motion retrieval, and reaches 68.94% R@1 and 82.86% R@2 in the motion-to-text retrieval, consistently surpassing LaMP (Li et al. 2025) and TMR(Li et al. 2025). On KIT-ML, our approach further pushes performance to 91.19% R@5 and 84.38% R@3, consistently surp assing baselines. We attribute these improvements to our noise augmentation strategy, which alleviates the limited motion diversity and text annotations in both datasets that lead to many hard negative samples, thereby enhancing the model's discriminative capability for subtle motion variations.

Plug-and-Play Capability of ReAlign. To demonstrate the plug-and-play capability and generalizability of our ReAlign, we integrate it into various diffusion-based models for text-to-motion generation, as shown in Table 4. Across methods such as Mo.Diffuse (Zhang et al. 2024), MDM (Tevet et al. 2023), MLD (Chen et al. 2023b), MotionLCM and its extension MLD++ (Dai et al. 2025). Our ReAlign consistently enhances performance. Notably, it achieves substantial improvements in alignment quality and motion realism, with relative gains of up to 18% in R@1 and 59% in FID for MLD. While diversity slightly decreases in some cases, this is expected and beneficial. Better diversity does not always indicate better quality, as it simply reflects motion variety. ReAlign prioritizes well-aligned motions over misaligned ones, leading to significant gains in other metrics without compromising generation quality. These results underscore the plug-and-play capability of this module, effectively elevating the efficacy of diverse motion generation frameworks.

Ablation Studies and Discussions

Effectiveness of Handling Noisy. To verify the necessity of handling noise and to avoid reward hacking, we varied the denoising steps of MLD from 1 to 50, employing both the reward model (RM) and the ReAlign to perform reward-guided sampling. This comples the model to generate noisy motions. As shown in Figure 4, compared to the baseline and the RM, ours achieves superior performance across all denoising steps, demonstrating that explicitly handling noise during the denoising benefits generating higher quality motions. Notably, we observed that MLD generates motions with certain semantics even in the early steps of denoising,

	Methods	Noise	Text-Motion Retrieval↑					Motion-Text Retrieval↑				
			R@1	R@2	R@3	R@5	R@10	R@1	R@2	R@3	R@5	R@10
HumanML3D	TEMOS (2022)	✗	40.49	53.52	61.14	70.96	84.15	39.96	53.49	61.79	72.40	85.89
	T2M (2022b)	✗	52.48	71.05	80.65	89.66	**96.58**	52.00	71.21	81.11	89.87	96.78
	TMR (2023)	✗	67.16	81.32	86.81	91.43	95.36	67.97	81.20	86.35	91.70	95.27
	LaMP (2025)	✗	67.18	81.90	87.04	**92.00**	95.73	68.02	82.10	87.50	92.20	**96.90**
	ReAlign (ours)	✓	**67.59**	**82.24**	**87.44**	91.97	96.28	**68.94**	**82.86**	**87.95**	**92.44**	96.28
KIT-ML	TEMOS (2022)	✗	43.88	58.25	67.00	74.00	84.75	41.88	55.88	65.62	75.25	85.75
	T2M (2022b)	✗	42.25	62.62	75.12	87.50	96.12	39.75	62.75	73.62	86.88	95.88
	TMR (2023)	✗	49.25	69.75	78.25	87.88	95.00	50.12	67.12	76.88	88.88	94.75
	ReAlign (ours)	✓	**52.84**	**71.66**	**82.96**	**91.19**	**97.59**	**52.98**	**72.87**	**84.38**	**92.61**	**96.87**

Table 3: Comparison of Text-to-motion (**left**) and motion-to-text (**right**) retrieval methods on the HumanML3D and KIT-ML datasets. The column "Noise" indicates whether the method can handle noisy motion from the denoised process.

Method	R Precision ↑			FID ↓	MM Dist ↓	Div. →
	Top 1	Top 2	Top 3			
Real	0.511	0.703	0.797	0.002	2.974	9.503
MDiff (2024)	0.491	0.681	0.775	0.630	3.113	9.410
w/ ReAlign	$0.534_{+9\%}$	$0.733_{+8\%}$	$0.829_{+7\%}$	$0.370_{+41\%}$	$2.807_{+10\%}$	$9.372_{-0.04}$
MDM (2023)	0.455	0.645	0.749	0.489	3.330	9.920
w/ ReAlign	$0.470_{+3\%}$	$0.677_{+5\%}$	$0.789_{+5\%}$	$0.325_{+34\%}$	$3.129_{+6\%}$	$9.355_{+0.27}$
MLD (2023b)	0.481	0.673	0.772	0.473	3.196	9.724
w/ ReAlign	$0.567_{+18\%}$	$0.759_{+13\%}$	$0.848_{+10\%}$	$0.195_{+59\%}$	$2.704_{+15\%}$	$9.474_{+0.19}$
MLCM[4] (2025)	0.502	0.698	0.798	0.304	3.012	9.607
w/ ReAlign	$0.540_{+8\%}$	$0.739_{+6\%}$	$0.833_{+4\%}$	$0.273_{+10\%}$	$2.797_{+7\%}$	$9.683_{-0.08}$
MLD++ (2025)	0.548	0.738	0.829	0.073	2.810	9.658
w/ ReAlign	$0.572_{+4\%}$	$0.764_{+4\%}$	$0.852_{+3\%}$	$0.055_{+25\%}$	$2.648_{+6\%}$	$9.478_{+0.13}$

Table 4: Performance improvement of motion generation methods with our step-aware reward guidance. Results are reported on the HumanML3D dataset, showing improvements over baseline methods. Div. stands for Diversity.

rather than purely noise. This behavior may be attributed to the latent VAE used in MLD, which exhibits robustness to noise. These experimental results show the feasibility and necessity of handling noise in denoising, aligning with conclusions drawn in DALLE-2 (2022) and GLIDE (2021).

Effectiveness of Reward Sampling. We assess the ReAlign, including T2M and M2M alignment rewards, along with the step-aware strategy in text-to-motion generation on MLD (Chen et al. 2023b). As shown in Table 5, results indicate that the T2M reward significantly improves the alignment between the generated motions and text descriptions, as well as the realism of the motions. While the M2M reward alone exhibits limited efficacy due to the inaccuracy of text-to-motion retrieval, its integration with the step-aware strategy further enhances motion realism, validating the necessity of handling noise during the sampling. The combination of T2M and step-aware strategies achieves optimal performance, with M2M providing additional realism gains.

Discussion on ReAlign and Classifier-free Guidance. As shown in Table 6, our ReAlign is compatible with CFG (Ho and Salimans 2022), and their integration can further unleash the performance of the diffusion model. Unlike CFG, which requires training, our ReAlign is plug-and-play and

T2M	M2M	SA	R Precision ↑			FID ↓	MM Dist ↓	Div. →
			Top 1	Top 2	Top 3			
✗	✗	✗	0.481	0.673	0.772	0.473	3.196	9.724
✓	✗	✗	0.556	0.747	0.841	0.213	2.761	**9.516**
✗	✓	✗	0.517	0.721	0.809	0.205	2.932	9.455
✓	✓	✗	0.556	0.750	0.840	0.199	2.750	9.529
✓	✗	✓	**0.568**	**0.761**	**0.850**	0.212	2.714	9.598
✗	✓	✓	0.523	0.709	0.810	0.203	2.963	9.525
✓	✓	✓	0.567	0.759	0.848	**0.195**	**2.704**	9.474

Table 5: Ablation study of the text-to-motion on HumanML3D dataset. "T2M", "M2M" and "SA" denote the text-to-motion reward, motion-to-motion reward and step-aware training, respectively, Div. stands for Diversity.

CFG	ReAlign	R Precision ↑			FID ↓	MM Dist ↓	Diversity →
		Top 1	Top 2	Top 3			
✗	✗	0.263	0.407	0.506	0.586	4.823	8.613
✓	✗	0.481	0.673	0.772	0.473	3.196	9.724
✓	✓	**0.567**	**0.759**	**0.848**	**0.195**	**2.704**	**9.474**

Table 6: Ablation study of the guidance strategy. Evaluation conducted on the HumanML3D with MLD (2023b) as the baseline. "CFG" denotes the classifier-free guidance.

supports flexible reward design for other tasks (e.g., physical reward, trajectory reward, style reward, etc.). In this work, we mainly focus on improving text-motion alignment, while future research will explore reward designs for more tasks.

Conclusion

We propose ReAlign, a plug-and-play reward-guided sampling strategy for diffusion-based text-to-motion generation. By jointly optimizing text-aligned and motion-aligned rewards during denoising, ReAlign effectively improves semantic consistency and motion realism. Our method integrates seamlessly with existing diffusion models without extra fine-tuning. Extensive experiments demonstrate that ReAlign achieves significant gains in both text-motion alignment and motion quality over state-of-the-art baselines.

Acknowledgments

This work is supported by National Natural Science Foundation of China (62302093, 62276134, 52441503), Jiangsu Province Natural Science Fund (BK20230833), Double First-Class Construction Foundation of China (23GH020227), the Fundamental Research Funds for the Central Universities (2242025K30024), the Open Research Fund of the State Key Laboratory of Multimodal Artificial Intelligence Systems (E5SP060116), and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Proposal ID: 23-SIS-SMU-070). We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

Chen, H.; Lu, C.; Wang, Z.; Su, H.; and Zhu, J. 2023a. Score regularized policy optimization through diffusion behavior. *arXiv preprint arXiv:2310.07297*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023b. Executing your Commands via Motion Diffusion in Latent Space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Dai, W.; Chen, L.-H.; Wang, J.; Liu, J.; Dai, B.; and Tang, Y. 2025. Motionlcm: Real-time controllable motion generation via latent consistency model. In *ECCV*, 390–408.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dou, Z.; and Song, Y. 2024. Diffusion posterior sampling for linear inverse problem solving: A filtering perspective. In *International Conference on Learning Representations*.

Guo, C.; Mu, Y.; Javed, M. G.; Wang, S.; and Cheng, L. 2024. Momask: Generative masked modeling of 3d human motions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1900–1910.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating Diverse and Natural 3D Human Motions from Text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5142–5151.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, 580–597. Springer.

Han, G.; Liang, M.; Tang, J.; Cheng, Y.; Liu, W.; and Huang, S. 2024. Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. *arXiv preprint arXiv:2410.07296*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Huang, Y.; Wan, W.; Yang, Y.; Callison-Burch, C.; Yatskar, M.; and Liu, L. 2024. CoMo: Controllable Motion Generation Through Language Guided Pose Code Editing. In *European Conference on Computer Vision*, 180–196. Springer-Verlag. ISBN 978-3-031-73396-3.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36: 20067–20079.

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*, 2151–2162.

Li, G.; Huang, Z.; and Wei, Y. 2025. Towards a mathematical theory for consistency training in diffusion models. In *International Conference on Artificial Intelligence and Statistics*, 1621–1629. PMLR.

Li, Z.; Yuan, W.; Qiu, L.; Zhu, S.; Gu, X.; Shen, W.; Dong, Y.; Dong, Z.; Yang, L. T.; et al. 2025. LaMP: Language-Motion Pretraining for Motion Generation, Retrieval, and Captioning. In *International Conference on Learning Representations*.

Liang, H.; Bao, J.; Zhang, R.; Ren, S.; Xu, Y.; Yang, S.; Chen, X.; Yu, J.; and Xu, L. 2024. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 482–493.

Liang, Z.; Yuan, Y.; Gu, S.; Chen, B.; Hang, T.; Cheng, M.; Li, J.; and Zheng, L. 2025. Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13199–13208.

Liu, X.; Mao, Y.; Zhou, W.; and Li, H. 2024. MotionRL: Align Text-to-Motion Generation to Human Preferences with Multi-Reward Reinforcement Learning. *arXiv preprint arXiv:2410.06513*.

Liu, X.; Park, D. H.; Azadi, S.; Zhang, G.; Chopikyan, A.; Hu, Y.; Shi, H.; Rohrbach, A.; and Darrell, T. 2023. More control for free! image synthesis with semantic diffusion guidance. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 289–299.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Meng, Z.; Xie, Y.; Peng, X.; Han, Z.; and Jiang, H. 2024. Rethinking diffusion for text-driven human motion generation. *arXiv preprint arXiv:2411.16575*.

Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; and Chen, M. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

Nisonoff, H.; Xiong, J.; Allenspach, S.; and Listgarten, J. 2025. Unlocking Guidance for Discrete State-Space Diffusion and Flow Models. In *International Conference on Learning Representations*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 480–497. Springer.

Petrovich, M.; Black, M. J.; and Varol, G. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *International Conference on Computer Vision*.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The kit motion-language dataset. *Big data*, 4(4): 236–252.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2): 3.

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*, 13756–13766.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*.

Tan, X.; Wang, H.; Geng, X.; and Zhou, P. 2025. SoPo: Text-to-Motion Generation Using Semi-Online Preference Optimization. *Advances in Neural Information Processing Systems*.

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*.

Uehara, M.; Zhao, Y.; Wang, C.; Li, X.; Regev, A.; Levine, S.; and Biancalani, T. 2025. Reward-guided controlled generation for inference-time alignment in diffusion models: Tutorial and review. *arXiv preprint arXiv:2501.09685*.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion model alignment using direct preference optimization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8228–8238.

Wan, Z.; Feng, X.; Wen, M.; Mcaleer, S. M.; Wen, Y.; Zhang, W.; and Wang, J. 2024. AlphaZero-Like Tree-Search can Guide Large Language Model Decoding and Training. In *International Conference on Machine Learning*, 49890–49920. PMLR.

Wang, H.; Weng, W.; Wang, J.; Zhao, F.; Xie, G.-S.; Geng, X.; and Wang, L. 2025. Foundation model for skeleton-based human action understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wu, B.; Xie, J.; Ding, M.; Kong, Z.; Ren, J.; Bai, R.; Qu, R.; and Shen, L. 2025a. FineMotion: A Dataset and Benchmark with both Spatial and Temporal Annotation for Fine-grained Motion Generation and Editing. *arXiv preprint arXiv:2507.19850*.

Wu, B.; Xie, J.; Shen, K.; Kong, Z.; Ren, J.; Bai, R.; Qu, R.; and Shen, L. 2025b. MG-MotionLLM: A unified framework for motion comprehension and generation across multiple granularities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 27849–27858.

Wu, Y.; Zhou, S.; Yang, M.; Wang, L.; Chang, H.; Zhu, W.; Hu, X.; Zhou, X.; and Yang, X. 2025c. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 8496–8504.

Yang, X.; Wu, Y.; Yang, M.; Chen, H.; and Geng, X. 2023. Exploring diverse in-context configurations for image captioning. *Advances in Neural Information Processing Systems*, 36: 40924–40943.

Yuan, W.; He, Y.; Shen, W.; Dong, Y.; Gu, X.; Dong, Z.; Bo, L.; and Huang, Q. 2025. Mogents: Motion generation based on spatial-temporal joint modeling. *Advances in Neural Information Processing Systems*, 37: 130739–130763.

Zhang, J.; Fan, H.; and Yang, Y. 2025. Energymogen: Compositional human motion generation with energy-based diffusion model in latent space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Zhang, J.; Zhang, Y.; Cun, X.; Zhang, Y.; Zhao, H.; Lu, H.; Shen, X.; and Shan, Y. 2023a. Generating Human Motion From Textual Descriptions With Discrete Representations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14730–14740.

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.

Zhang, M.; Guo, X.; Pan, L.; Cai, Z.; Hong, F.; Li, H.; Yang, L.; and Liu, Z. 2023b. Remodiffuse: Retrieval-augmented motion diffusion model. In *IEEE/CVF International Conference on Computer Vision*, 364–373.

Zhang, M.; Jin, D.; Gu, C.; Hong, F.; Cai, Z.; Huang, J.; Zhang, C.; Guo, X.; Yang, L.; He, Y.; et al. 2025a. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*. Springer.

Zhang, M.; Li, H.; Cai, Z.; Ren, J.; Yang, L.; and Liu, Z. 2023c. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36: 13981–13992.

Zhang, Z.; Liu, A.; Reid, I.; Hartley, R.; Zhuang, B.; and Tang, H. 2025b. Motion Mamba: Efficient and Long Sequence Motion Generation. In *European Conference on Computer Vision*, 265–282. Springer.

Zou, Q.; Yuan, S.; Du, S.; Wang, Y.; Liu, C.; Xu, Y.; Chen, J.; and Ji, X. 2025. ParCo: Part-Coordinating Text-to-Motion Synthesis. In *European Conference on Computer Vision*, 126–143.