Bilingual Text-to-Motion Generation via Step-Aware Reward-Guided Alignment

Anonymous ICCV submission

Paper ID 2603

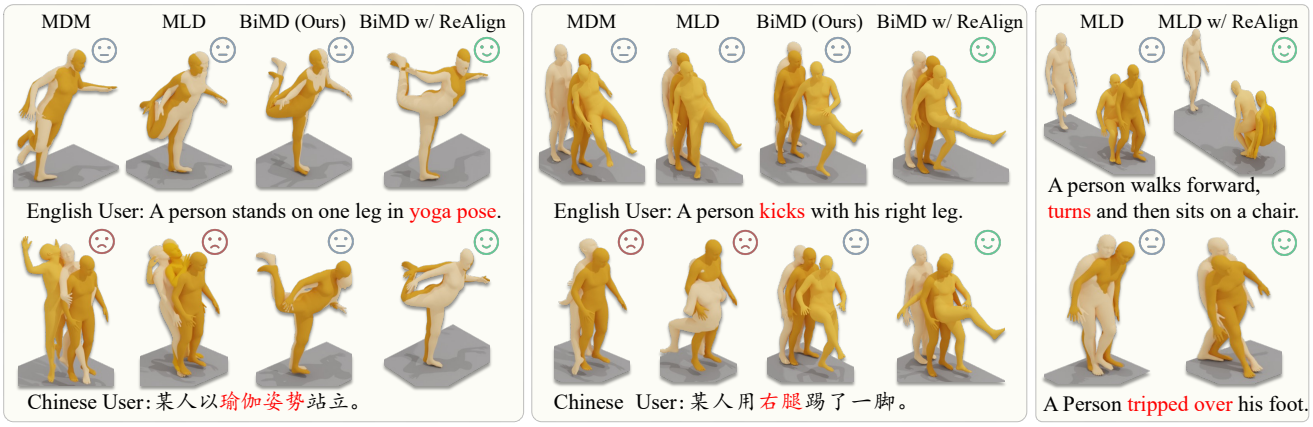

Figure 1. Visual results of bilingual (left and middle) and monolingual (right) text-driven motion generation. This figure presents motions generated by existing methods, such as MDM [27] and MLD [2], alongside our Bilingual Motion Diffusion model (BiMD) and plug-and-play Reward-guided Alignment (ReAlign), across both tasks. Observations reveal that: (1) in bilingual motion generation, MDM and MLD exhibit limitations in processing bilingual inputs; (2) in monolingual motion generation, generated motions demonstrate persistent misalignment with input texts. The left figure shows that our BiMD successfully generates motion from both English and Chinese inputs. Furthermore, the right figure highlights that BiMD, integrated with our ReAlign, successfully mitigates the misalignment issue.

Abstract

*Bilingual text-to-motion generation, which synthesizes 3D human motions from bilingual text inputs, holds immense potential for cross-linguistic applications in gaming, film, and robotics. However, this task faces critical challenges: the absence of bilingual motion-language datasets and the misalignment between text and motion distributions in diffusion models, leading to semantically inconsistent or low-quality motions. To address these challenges, we propose BiHumanML3D, a novel bilingual human motion dataset, which establishes a crucial benchmark for bilingual text-to-motion generation models. Furthermore, we propose a **Bi**lingual **M**otion **D**iffusion model (**BiMD**), which leverages cross-lingual aligned representations to capture semantics, thereby achieving a unified bilingual model. Building upon this, we propose **Re**ward-guided sampling **Align**ment (**ReAlign**) method, comprising a step-aware reward model to assess alignment quality during 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. Experiments demonstrate that our approach significantly improves text-motion alignment and motion quality compared to existing state-of-the-art methods.*

1. Introduction

With the increasing demand for realistic and diverse 3D motion in gaming, filmmaking, and robotics [2, 6, 7, 12], text-to-motion generation has emerged as a key research topic, offering intuitive text-based control. Particularly, bilingual text-to-motion generation which aims to generate motion from bilingual text descriptions, holds significant potential in cross-linguistic applications [14, 41]. However, this task remains largely unexplored due to several key challenges.

Firstly, the scarcity of bilingual text-motion datasets presents a significant challenge to the development of bilingual text-to-motion generation. While many large-scale English motion datasets [7, 20] have been established and widely utilized, there are no publicly available motion

datasets annotated in other languages, let alone bilingual datasets. This data scarcity greatly restricts research and progress in bilingual text-to-motion generation methods.

Secondly, diffusion-based models [27, 34] struggle with text-motion alignment due to their reliance on text embeddings encoded by CLIP [22] 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, resulting in synthesized motions that lack coherence with the input descriptions as shown in Fig. 1.

To the best of our knowledge, this is the first study to explore bilingual text-to-motion generation. While research in bilingual generative models is gaining traction, most efforts remain domain-specific. For instance, Zuo et al. [41] developed a bilingual 3D sign language avatar generator, and Li et al. [14] introduced a bilingual image diffusion transformer. However, these works are designed for specific applications, and do not generalize to motion synthesis, leaving bilingual text-to-motion generation an open challenge. Furthermore, prior methods aiming to improve text-to-motion alignment, such as reinforcement learning with reward functions [9, 16, 25], primarily focus on fine-tuning generative models to enhance motion quality. However, these approaches do not explicitly address misalignment between text and motion. In contrast, we propose a plug-and-play reward model that can be seamlessly integrated into any diffusion model without additional fine-tuning, ensuring enhanced text-motion alignment.

Contributions. Our first contribution lies in the introduction of a pioneering bilingual text-to-motion dataset, BiHumanML3D, accompanied by a corresponding bilingual text-to-motion method, **Bi**lingual **M**otion **D**iffusion (**BiMD**). To address the scarcity of bilingual text-motion datasets, we extend the widely used text-to-motion dataset, HumanML3D [7], by introducing its bilingual version, BiHumanML3D. Specifically, a multi-stage translation pipeline based on large language models and manual correction is designed to ensure high-quality annotations and accurate semantic translations. Furthermore, a unified bilingual motion diffusion model, BiMD, is trained to efficiently handle bilingual text-to-motion generation. This is enabled by harmonizing semantics across languages via cross-lingual alignment.

Additionally, 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. To this end, 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.

Finally, extensive experiments show that BiMD, driven by reward-guided sampling, generates high-quality motions aligned with both English and Chinese semantics. The step-aware reward model enhances diffusion-based models with plug-and-play versatility. On BiHumanML3D, BiMD with cross-lingual alignment outperforms baselines, including language-specific models, proving its strength in bilingual motion generation. On monolingual HumanML3D, our reward model enhances both our BiMD and the previous SoTA MLD++ [6], e.g., improving BiMD by **55.2%** and MLD++ by **24.7%** in terms of FID, without any additional training. These results highlight the superiority of our BiMD and ReAlign strategy.

2. Related Works

Text-to-Motion Generation. Text-to-motion generation represents a critical task in computer vision, exhibiting rapid advancements in recent years [26, 30, 32, 33, 36, 39]. Specifically, Tevet et al.[27] and Zhang et al.[34] first proposed diffusion models to address text-driven motion generation, laying the groundwork for the following innovations. Subsequently, Dai et al. [6] presented MotionLCM, a real-time controllable model that refines motion-latent diffusion, enabling precise spatiotemporal control via few-step inference. Zhang et al.[38] introduced motion mamba, a state-space framework that leverages hierarchical temporal and bidirectional spatial modules to improve efficiency and long-sequence modeling. However, these works are still confined to monolingual settings. Bilingual text-to-motion generation remains a challenging area due to the lack of datasets and misalignment between different languages.

Alignment in Motion Generation. Alignment represents a versatile technique widely employed across the domains of language modeling [23], image generation [28], and policy optimization [1]. Recently, researchers [9, 16, 25] have explored human preference alignment in text-to-motion generation. Han et al. [9] introduced ReinDiffuse, a diffusion-based model refined through reinforcement learning, to enhance the physical plausibility of generated motions. Liu et al. [16] investigated aligning human preferences using the proposed multi-reward reinforcement learning framework. Tan et al. [25] proposed SoPo, a novel semi-online preference optimization method that refines text-to-motion models. 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.

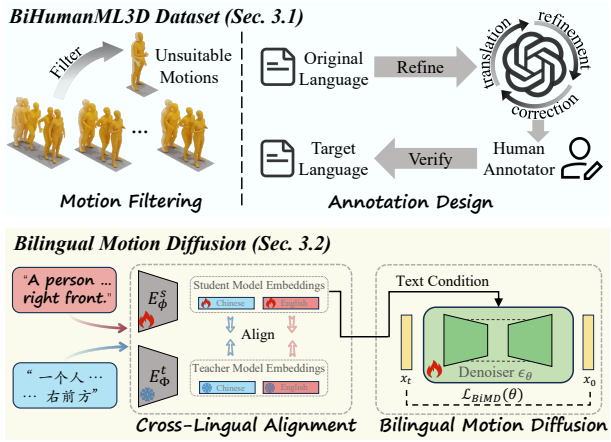

Figure 2. Pipeline for constructing our bilingual HumanML3D dataset (top) and training the bilingual motion diffusion model (bottom). We align English and Chinese text semantics in a shared latent space by freezing the teacher model E^t_Φ and fine-tuning the student model E^s_ϕ with the cross-lingual alignment loss $\mathcal{L}_{CLA}$ in Eq. (1). The aligned student model E^s_ϕ is then provided text conditions for training the diffusion model ϵ_θ, enabling bilingual motion generation while minimizing $\mathcal{L}_{\text{BiMD}}$ in Eq. (2).

3. BiHumanML3D Dataset & Bilingual Model

We address the challenges of bilingual motion generation by introducing a new dataset and a unified bilingual motion diffusion model. In Sec. 3.1, we present a bilingual dataset to bridge the gap in multilingual motion generation. Instead of training separate models for each language—an inefficient approach that fails to leverage shared semantics—we propose a cross-lingually aligned diffusion model in Sec. 3.2, ensuring consistent motion generation across languages. Lastly, to tackle misalignment between text and motion, we introduce reward-guided sampling in Sec. 4, refining motion quality and semantic coherence.

3.1. Bilingual HumanML3D Dataset

Despite the progress in text-to-motion models, their reliance on English-only datasets limits their usability in multilingual applications, reducing accessibility for non-English speakers and restricting cultural diversity in global industries like animation and robotics. So we construct a bilingual motion dataset for multilingual motion synthesis via two stages. See more construction details in Appendix B.

Data Collection & Filtering. Since large-scale English text-to-motion datasets are already well-established, we extend HumanML3D [7] into a bilingual version. Following previous work [2, 27], we filter out excessively short or long motion clips which often lack meaningful semantic signals, ensuring high-quality motion-caption pairs for annotation.

Annotation Design. To generate bilingual annotations while preserving motion semantics, we develop an LLM-assisted translation pipeline as shown in Fig. 2. This process involves three stages: (1) initial translation of English captions into Chinese using an LLM, (2) refinement to enhance translation quality, and (3) validation via both automated and human review to ensure linguistic accuracy.

By introducing the Bilingual HumanML3D Dataset that contains 13,312 bilingual motions, we take a crucial step toward removing language barriers in text-to-motion generation. This dataset enhances the adaptability of motion synthesis models across languages while promoting fairness and inclusivity in AI-driven animation and robotics.

3.2. Bilingual Motion Diffusion

While training separate diffusion models for each language on our bilingual dataset is a straightforward approach, it is computationally expensive and overlooks semantic similarities between languages. To address this, we propose a unified bilingual motion diffusion model that leverages cross-lingual alignment and bilingual diffusion training to reduce costs and enhance motion diversity and quality across languages, as illustrated in Fig. 2.

Cross-Lingual Alignment. Inspired by AltCLIP [3], we enhance cross-lingual text understanding by fine-tuning the pretrained language model XLM [5] using knowledge distillation. The goal is to align sentence semantics across languages, ensuring that motion descriptions in different languages share a consistent latent representation, allowing the motion diffusion model to interpret them interchangeably. Specifically, we align the text embeddings of a student model E^s_ϕ from XLM [5] with those of a teacher model E^t_Φ from OpenCLIP [4] by optimizing:

$$\mathcal{L}_{\text{CLA}}(\phi) = \mathrm{D}_{\text{KL}}(\mathrm{F}^{\Phi}_{\text{en}}|\mathrm{F}^{\phi}_{\text{cn}}) + \mathrm{D}_{\text{KL}}(\mathrm{F}^{\phi}_{\text{cn}}|\mathrm{F}^{\Phi}_{\text{en}}), \quad (1)$$

where F^M_{en} and F^M_{cn} represent the English and Chinese text embeddings encoded by model M. Minimizing this loss ensures that the student model E^s_ϕ learns to interpret text descriptions in both languages interchangeably. This alignment is crucial for maintaining motion fidelity and coherence without relying on separate models for each language.

Bilingual Training. For robust bilingual motion generation, we use a motion diffusion model [2] as our generative backbone, conditioning it on cross-lingually aligned text embeddings. Instead of training two distinct models, we introduce a language-agnostic training strategy: each motion sample is randomly paired with either an English or Chinese text description. This encourages the model to capture shared motion patterns while remaining sensitive to language-specific nuances. Our training loss is as:

$$\mathcal{L}_{\text{BiMD}}(\theta) = \mathbb{E}_{\epsilon,t,c}\left[\left\|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, t, E^s_\phi\left(c_s\right)\right)\right\|_2^2\right], \quad (2)$$

where c_s is the randomly selected English or Chinese description of motion $\mathbf{x}$, and $E^s_\phi(c_s)$ represents its cross-

lingually aligned feature. This approach enables the model to synthesize high-quality motion sequences from multilingual text prompts without requiring separate models.

By integrating cross-lingual alignment with bilingual diffusion training, our method overcomes the English-only limitation of existing text-to-motion models, making text-driven motion generation more accessible and effective across languages. This paves the way for broader applications like virtual animation and multilingual interactions.

4. Step-Aware Reward-Guided Alignment

4.1. Motivation & Framework

Preliminaries. Existing diffusion-based motion generation methods [2, 27] 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) [24]:

$$\mathrm{d}\mathbf{x} = \mathbf{f}(\mathbf{x}, t)\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{3}$$

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 [24]:

$$\mathrm{d}\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla \log p_t(\mathbf{x})]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}, \tag{4}$$

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

Motivation. While bilingual diffusion models enable motion generation across languages, they often fail to produce motions that accurately align with textual descriptions. For example, as illustrated in Fig. 3, 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 (Fig. 3), 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 [22] 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 retraining. As shown in Fig. 3, sampling from this ideal distribution ensures both high-probability density and strong semantic alignment, overcoming previous limitations.

Overall 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. Sec. 4.2 details the estimation of the reward distribution, while Sec. 4.3 describes the motion sampling process.

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

$$p_t^I(\mathbf{x}|c) = p_t(\mathbf{x}|c)p_t^r(\mathbf{x}|c)/Z(c), \tag{5}$$

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. (5) to replace the vanilla sampling distribution $p_t(\mathbf{x}|c)$, the reverse SDE becomes:*

$$\mathbf{d}\mathbf{x} = \Big[\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla\big(\log p_t(\mathbf{x}|c) + \log p_t^r(\mathbf{x}|c)\big)\Big]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{6}$$

See its proof in App. D.1. 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 (Sec. 4.2) and outline the motion sampling procedure (Sec. 4.3).

4.2. 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 diffusion process. Existing methods [19] assume clean and noise-free motion sequences, 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

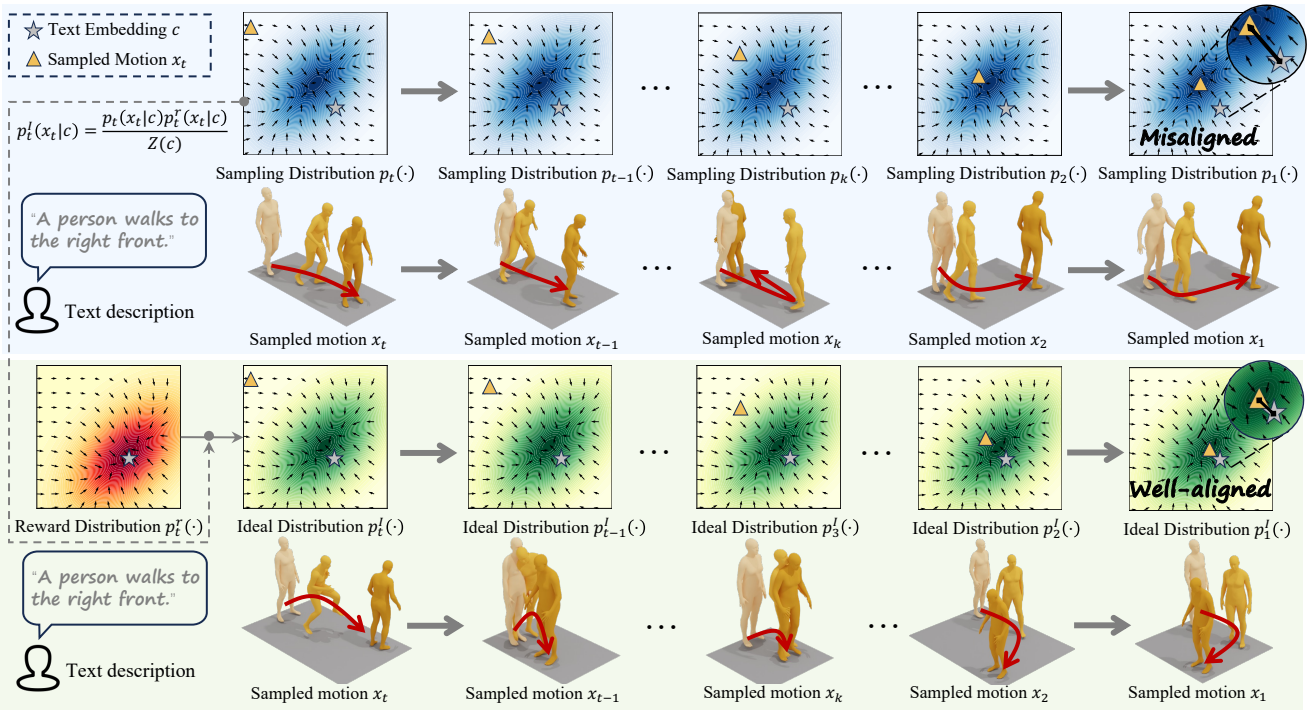

Figure 3. Illustration of the sampling process in diffusion-based motion generation frameworks. The blue region represents the sampling distribution $p_t(\cdot)$ learned by the diffusion model, while the green region depicts the ideal sampling distribution $p_t^I(\cdot)$ achieved by incorporating our proposed reward-guided sampling strategy with the sampling distribution $p_t(\cdot)$.

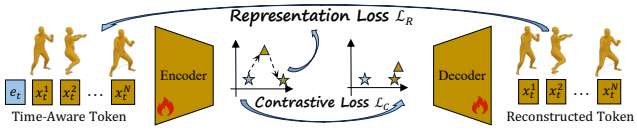

Figure 4. Framework of step-aware reward model. During this process, time-aware tokens, consisting of timestep embedding t and motion embeddings x_t^k, are aligned with text embedding c in the latent space and reconstructed via the decoder, with the encoder and decoder jointly optimized by contrastive loss $\mathcal{L}_C$ [19] and representation loss $\mathcal{L}_R$ [18].

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 Fig. 4, which explicitly accounts for noise variations across diffusion timesteps. Unlike conventional alignment models [19], 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 $\mathcal{L}_R$ [18] to learn meaningful motion embeddings, and a contrastive loss $\mathcal{L}_C$ [19] to ensure accurate text-motion retrieval:

$$\mathcal{L}_{RM}(\varphi; \mathbf{x}_t, c) = \mathcal{L}_C(\varphi; \mathbf{x}_t, c) + \mathcal{L}_R(\varphi; \mathbf{x}_t, c). \quad (7)$$

See more details to train this model in App. C.3.

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:

$$R_\varphi(\mathbf{x}, c) = \cos(\mathbf{z}_\mathbf{x}, \mathbf{z}_c), \quad (8)$$

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:

$$\mathbf{x}^c = \arg\max_{\mathbf{x}\in\mathcal{D}_{tr}} R_\varphi(\mathbf{x}, c). \tag{9}$$

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:

$$R_m(\mathbf{x}_t, c) = \cos(\mathbf{z}_{\mathbf{x}}, \mathbf{z}_{\mathbf{x}^c}), \tag{10}$$

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:

$$R(\mathbf{x}_t, c) = \mu R_\varphi(\mathbf{x}_t, c) + \eta R_m(\mathbf{x}_t, c), \tag{11}$$

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:

$$p_t^r(\mathbf{x}_t|c) = \exp\left(R(\mathbf{x}_t, c)\right)/Z^r(c). \tag{12}$$

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 approach constructs a robust reward signal that captures both semantic consistency and motion coherence. This enables more precise guidance of the diffusion sampling process, ensuring that generated motions are not only probable but also faithful to their textual descriptions.

4.3. 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. (12), the reverse SDE can be rewritten as:*

$$\mathrm{d}\mathbf{x} = \Big[\mathbf{f}(\mathbf{x}, t) - g(t)^2\nabla\big(\log p_t(\mathbf{x}|c) + R(\mathbf{x}_t, c)\big)\Big]\mathrm{d}t + g(t)\mathrm{d}\mathbf{w}. \tag{13}$$

See its proof in App. D.2. 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 [10] framework in the following theorem. See proof in App. D.4.

Algorithm 1 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. (15) to generate $\mathbf{x}_{t-1}$
end for
return $\mathbf{x}_0$

Theorem 3. *Given a reverse SDE defined in Eq. (13), adopting standard DDPM settings [10, 24] 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:*

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\bar{\mathbf{x}}_{t-1} + \sqrt{\beta_t}\epsilon\right) + \frac{\beta_t}{\sqrt{\alpha_t}}\nabla R(\mathbf{x}_t, c), \tag{14}$$

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 network, 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 (see detailed discussion in App. D.4), we remove the weight $\frac{\beta_t}{\sqrt{\alpha_t}}$ on the reward term, leading to a revised denoising process:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\bar{\mathbf{x}}_{t-1} + \sqrt{\beta_t}\epsilon\right) + \nabla R(\mathbf{x}_t, c). \tag{15}$$

Based on this theoretical framework, we introduce Algorithm 1 that integrates the step-aware reward model into the diffusion-based generation process.

5. Experiment

5.1. Experiment Setting

Datasets & Evaluation. We employ two widely used text-to-motion datasets, HumanML3D [7] and KIT-ML [21], alongside the proposed BiHumanML3D, for evaluation purposes. Consistent with the majority of prior studies [8, 13], 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. Our bilingual motion diffusion model adopts MLD [2] as the backbone, adhering to its

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 (2022) [7]	$0.455^{\pm0.002}$	$0.636^{\pm0.003}$	$0.736^{\pm0.003}$	$1.087^{\pm0.002}$	$3.347^{\pm0.008}$	$9.175^{\pm0.002}$
MDM (2023) [27]	$0.455^{\pm0.006}$	$0.645^{\pm0.007}$	$0.749^{\pm0.006}$	$0.489^{\pm0.047}$	$3.330^{\pm0.25}$	$9.920^{\pm0.083}$
T2M-GPT (2023) [31]	$0.492^{\pm0.003}$	$0.679^{\pm0.002}$	$0.775^{\pm0.002}$	$0.141^{\pm0.005}$	$3.121^{\pm0.009}$	$9.722^{\pm0.082}$
MLD (2023) [2]	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	$0.473^{\pm0.013}$	$3.196^{\pm0.010}$	$9.724^{\pm0.082}$
Mo.Diffuse (2024) [34]	$0.491^{\pm0.001}$	$0.681^{\pm0.001}$	$0.775^{\pm0.001}$	$0.630^{\pm0.001}$	$3.113^{\pm0.001}$	$9.410^{\pm0.49}$
OMG (2024) [15]	-	-	$0.784^{\pm0.002}$	$0.381^{\pm0.008}$	-	$9.657^{\pm0.085}$
MotionLCM (2025) [6]	$0.502^{\pm0.003}$	$0.698^{\pm0.002}$	$0.798^{\pm0.002}$	$0.304^{\pm0.012}$	$3.012^{\pm0.007}$	$9.607^{\pm0.066}$
LMM-T^2 (2024) [35]	$0.496^{\pm0.002}$	$0.685^{\pm0.002}$	$0.785^{\pm0.002}$	$0.415^{\pm0.002}$	$3.087^{\pm0.012}$	$9.176^{\pm0.074}$
Mo.Mamba (2025) [37]	$0.502^{\pm0.003}$	$0.693^{\pm0.002}$	$0.792^{\pm0.002}$	$0.281^{\pm0.011}$	$3.060^{\pm0.000}$	$9.871^{\pm0.084}$
CoMo (2024) [11]	$0.502^{\pm0.002}$	$0.692^{\pm0.007}$	$0.790^{\pm0.002}$	$0.262^{\pm0.004}$	$3.032^{\pm0.015}$	$9.936^{\pm0.066}$
ParCo (2025) [40]	$0.515^{\pm0.003}$	$0.706^{\pm0.003}$	$0.801^{\pm0.002}$	$0.109^{\pm0.005}$	$2.927^{\pm0.008}$	$9.576^{\pm0.088}$
BiMD (Ours)	$0.499^{\pm0.002}$	$0.691^{\pm0.002}$	$0.789^{\pm0.003}$	$0.397^{\pm0.011}$	$3.105^{\pm0.009}$	$9.635^{\pm0.089}$
w/ ReAlign (Ours)	$0.566^{\pm0.003}$ (+13.4%)	$0.759^{\pm0.002}$ (+9.7%)	$0.847^{\pm0.002}$ (+7.4%)	$0.178^{\pm0.006}$(+55.2%)	$2.714^{\pm0.007}$ (+12.6%)	$9.573^{\pm0.068}$ (+47.0%)
MLD++ [6] (Baseline)	$0.548^{\pm0.003}$	$0.738^{\pm0.003}$	$0.829^{\pm0.002}$	$0.073^{\pm0.003}$	$2.810^{\pm0.008}$	$9.658^{\pm0.089}$
w/ ReAlign (Ours)	**0.572**$^{\pm0.002}$ (+4.4%)	**0.764**$^{\pm0.002}$ (+3.5%)	**0.852**$^{\pm0.001}$ (+2.8%)	**0.055**$^{\pm0.003}$ (+24.7%)	**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.** The arrows ↑, ↓, and → indicate higher, lower, and closer-to-real-motion values are better, respectively. **Bold** highlights the best results. Percentages in subscripts indicate improvements over respective baselines. Our BiMD adopts a similar backbone of MLD's [2], and surpasses it on all metrics.

Method	R Precision ↑			FID ↓	MM Dist ↓	Diversity →
	Top 1	Top 2	Top 3			
Real	0.424	0.649	0.779	0.031	2.788	11.08
T2M (2022) [7]	0.361	0.559	0.681	3.022	2.052	10.72
MLD (2023) [2]	0.390	0.609	0.734	0.404	3.204	10.80
T2M-GPT (2023) [31]	0.416	0.627	0.745	0.514	3.007	10.86
CoMo (2024) [11]	0.422	0.638	0.765	0.332	2.873	10.95
Mo.Mamba (2025) [37]	0.419	0.645	0.765	0.307	3.021	11.02
ParCo (2025) [40]	0.430	0.649	0.772	0.453	2.820	10.95
Baseline [34]	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
Baseline (MDM) [27]	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. Since the models MLD [2] and MLD++ [6] for the KIT-ML dataset have not been released, we use the widely used MDM [27] as the baseline.

Method	CLA	Lang.	R Precision ↑			FID ↓	MM Dist ↓	Diversity →
			Top 1	Top 2	Top 3			
Real	-	CN	0.543	0.732	0.821	0.002	3.338	10.750
	-	EN	0.511	0.703	0.797	0.002	2.974	9.503
Mo.Diffuse [34]	✗	CN	0.478	0.680	0.783	1.024	3.512	11.586
	✓	CN	0.502	0.696	0.791	0.643	3.356	11.064
	✗	EN	0.491	0.681	0.782	0.630	3.113	9.410
MDM [27]	✗	CN	0.481	0.673	0.774	0.908	3.482	11.674
	✓	CN	0.497	0.693	0.791	0.627	3.347	11.612
	✗	EN	0.455	0.645	0.749	0.489	3.330	9.920
MLD [2]	✗	CN	0.482	0.671	0.769	0.789	3.557	11.204
	✗	EN	0.481	0.673	0.772	0.473	3.196	9.724
BiMD (Ours)	✓	CN	**0.505**	**0.696**	**0.792**	**0.528**	**3.338**	**10.741**
	✓	EN	**0.499**	**0.691**	**0.789**	**0.397**	**3.105**	**9.635**

Table 3. **Comparison of text-to-motion generation performance on the BiHumanML3D dataset.** "Lang." indicates the evaluated language. English and Chinese results are assessed from the original evaluator [7] and our proposed evaluator, respectively. **Bold** highlights the best results. Given the absence of bilingual motion generation methods in current literature, our study adapts established monolingual frameworks [2, 27, 34], training separate, language-specific models to advance the field. The symbol ✓" at CLA" denotes methods that employ our cross-lingual alignment representation, indicating the use of a unified model for generation, whereas the symbol "✗" means methods trained specifically for individual languages.

training configuration. For the cross-lingual alignment, we employ OpenCLIP [4] as the teacher model and use XLM-B [5] as the backbone of the student model, optimizing with $\mathcal{L}_{CLA}$ defined in Eq. (1). The step-aware reward model is built based on SkipTransformer [2]. The architecture consists of a transformer encoder for text and motion and a decoder for motion, all with 9 layers and 4 heads, with a latent space dimension of 256. During training, the probability of noisy motion augmentation is set to 0.5, and the model is aware of a maximum timestep of 1000. We use AdamW [17] as the optimizer with a learning rate of 10^{-4}, while other hyperparameter settings follow the TMR [19]. More details about experiment settings and results are provided in App. C.

5.2. Main Results

Text-to-Motion Generation. As shown in Tab. 1, our reward-guided sampling significantly enhances performance when integrated with state-of-the-art text-to-motion models. Specifically, our BiMD achieves performance comparable to previous SoTA methods and surpasses them by integrating the proposed reward-guided sampling. Notably, by using our proposed reward-guided sampling, MLD++ [6] achieves new SoTA results, with an R Precision@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 reward-guided sampling significantly enhances the performance of MDM [2], yielding SoTA results on the KIT-ML dataset, with an R Precision@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 reward-guided sampling demon-

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
MDiff [34]	0.491	0.681	0.775	0.630	3.113	9.410
w/ ReAlign	$0.534_{+8.8\%}$	$0.733_{+7.6\%}$	$0.829_{+7.0\%}$	$0.370_{+41.3\%}$	$2.807_{+9.9\%}$	$9.372^{-0.04}_{-40.9\%}$
MDM [27]	0.455	0.645	0.749	0.489	3.330	9.920
w/ ReAlign	$0.470_{+3.3\%}$	$0.677_{+5.0\%}$	$0.789_{+5.3\%}$	$0.325_{+33.5\%}$	$3.129_{+6.0\%}$	$9.355^{+0.27}_{+64.5\%}$
MLD [2]	0.481	0.673	0.772	0.473	3.196	9.724
w/ ReAlign	$0.567_{+17.9\%}$	$0.759_{+12.8\%}$	$0.848_{+9.8\%}$	$0.195_{+58.8\%}$	$2.704_{+15.4\%}$	$9.474^{+0.19}_{+86.9\%}$
MLCM[1] [6]	0.546	0.743	0.837	0.072	2.767	9.577
w/ ReAlign	$0.555_{+1.7\%}$	$0.751_{+1.1\%}$	$0.841_{+0.5\%}$	$0.088_{-22.2\%}$	$2.726_{+1.5\%}$	$9.541^{+0.04}_{+48.6\%}$
MLCM[4] [6]	0.502	0.698	0.798	0.304	3.012	9.607
w/ ReAlign	$0.540_{+7.6\%}$	$0.739_{+5.9\%}$	$0.833_{+4.4\%}$	$0.273_{+10.2\%}$	$2.797_{+7.1\%}$	$9.683^{-0.08}_{-73.1\%}$
MLD++ [6]	0.548	0.738	0.829	0.073	2.810	9.658
w/ ReAlign	$0.572_{+4.4\%}$	$0.764_{+3.5\%}$	$0.852_{+2.8\%}$	$0.055_{+24.7\%}$	$2.648_{+5.8\%}$	$9.478^{+0.13}_{+83.9\%}$
BiMD	0.499	0.691	0.789	0.397	3.105	9.635
w/ ReAlign	$0.566_{+13.4\%}$	$0.759_{+9.7\%}$	$0.847_{+7.4\%}$	$0.178_{+55.2\%}$	$2.714_{+12.6\%}$	$9.573^{+0.06}_{+47.0\%}$

Table 4. **Performance enhancement of motion generation methods with plug-and-play step-aware reward guidance.** Results are evaluated on the HumanML3D dataset, with improvements reported relative to baseline methods. Here, MLCM[1] and MLCM[4] denote the 1-step and 4-step model in MotionLCM [6]. MDiff is an abbreviation of MotionDiffuse [34].

strate the effectiveness of our reward-guided sampling in enhancing text-motion alignment quality.

Bilingual Text-to-Motion Generation. Table 3 reports bilingual text-to-motion generation results on our BiHumanML3D dataset. For a fair evaluation, we adapt state-of-the-art models, Mo.Diffuse, MLD, and MDM, to the Chinese context using the corresponding BiHumanML3D subset, comparing our cross-lingual alignment (CLA) approach against a language-specific baseline. Since these models are designed for English text-to-motion generation, we use ChineseCLIP [29] to replace their text encoder, adapting to Chinese settings. To extend these models for bilingual settings, we further substitute ChineseCLIP [29] with our CLA encoder. For all baselines, CLA integration yields notable gains; for example, in MDM, R-Precision rises from 77.4% to 79.1%, and FID drops from 0.908 to 0.627 compared to non-CLA models. This underscores the critical role of cross-lingual alignment in bilingual text-to-motion generation. Moreover, our BiMD model surpasses all baselines, achieving superior performance in both languages. These results highlight the synergy between BiMD and CLA, as well as the substantial advancement our approach offers over existing state-of-the-art methods.

Plug-and-Play Functionality of ReAlign. To demonstrate the plug-and-play capability and generalizability of our step-aware reward-guided alignment, we integrate it into various baseline models for text-to-motion generation, as shown in Tab. 4. Across methods such as Mo.Diffuse [34], MDM [27], MLD [2], MotionLCM [6], MLD++ [6], and our BiMD, our ReAlign consistently enhances performance. Notably, it achieves substantial improvements in alignment quality and motion realism, with relative gains of up to 17.9% in R Precision Top 1 and 66.1% in FID for BiMD. The approach also refines multimodal alignment while maintaining diversity close to real motions. 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 versatility of this module, effectively elevating the efficacy of diverse motion generation frameworks.

T2M	M2M	SA	R Precision ↑			FID ↓	MM Dist ↓	Diversity →
			Top 1	Top 2	Top 3			
✗	✗	✗	0.499	0.691	0.789	0.397	3.105	9.635
✓	✗	✗	0.557	0.749	0.840	0.216	2.760	**9.513**
✗	✓	✗	0.522	0.715	0.809	0.188	2.932	0.455
✓	✓	✗	0.557	0.751	0.841	0.182	2.748	9.530
✓	✗	✓	**0.567**	**0.760**	**0.849**	0.196	2.721	9.598
✗	✓	✓	0.521	0.711	0.806	0.210	2.961	9.526
✓	✓	✓	0.566	0.759	0.847	**0.178**	**2.714**	9.573

Table 5. **Ablation study of the text-to-motion on HumanML3D dataset.** Here, "T2M", "M2M" and "SA" denote whether the text-to-motion reward, motion-to-motion reward and step-aware training is used, respectively.

5.3. Ablation Study

Effectiveness of Reward Model. We assess the dual reward mechanism, including T2M and M2M alignment rewards, along with the step-aware strategy in text-to-motion generation. As shown in Tab. 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 has limited impact, combining it with the step-aware strategy further enhances motion realism. The combination of T2M and step-aware already yields the best result, with the addition of M2M providing a further boost to realism.

6. Conclusion

This paper introduces bilingual text-to-motion generation with BiHumanML3D, the first bilingual dataset, and BiMD, a unified diffusion model utilizing cross-lingual alignment for efficient motion generation. To tackle text-motion misalignment, we propose a step-aware reward model plugged into a pretrained diffusion model, enabling reward-guided sampling without further training.

Limitation discussion. Our BiMD aims to sample motion from the pretrained motion representation space. However, the scarcity of motion data limits the ability of pretrained models, such as variational autoencoders, to effectively extract rich semantic features, thereby constraining the generalizability of our model.

References

[1] Huayu Chen, Cheng Lu, Zhengyi Wang, Hang Su, and Jun Zhu. Score regularized policy optimization through diffusion behavior. *arXiv preprint arXiv:2310.07297*, 2023. 2

[2] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 1, 3, 4, 6, 7, 8

[3] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*, 2022. 3

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 3, 7

[5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 3, 7

[6] Wenxun Dai, Ling-Hao Chen, Jingbo Wang, Jinpeng Liu, Bo Dai, and Yansong Tang. Motionlcm: Real-time controllable motion generation via latent consistency model. In *European Conference on Computer Vision*, pages 390–408, Cham, 2024. Springer Nature Switzerland. 1, 2, 7, 8

[7] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5142–5151, 2022. 1, 2, 3, 6, 7

[8] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1900–1910, 2024. 6

[9] Gaoge Han, Mingjiang Liang, Jinglei Tang, Yongkang Cheng, Wei Liu, and Shaoli Huang. Reindiffuse: Crafting physically plausible motions with reinforced diffusion model. *arXiv preprint arXiv:2410.07296*, 2024. 2

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 6

[11] Yiming Huang, Weilin Wan, Yue Yang, Chris Callison-Burch, Mark Yatskar, and Lingjie Liu. Como: Controllable motion generation through language guided pose code editing. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIX*, page 180–196, Berlin, Heidelberg, 2024. Springer-Verlag. 7

[12] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[13] Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, Xiaodong Gu, Weichao Shen, Yuan Dong, Zilong Dong, and Laurence T. Yang. Lamp: Language-motion pretraining for motion generation, retrieval, and captioning. 2024. 6

[14] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchi Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al. Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024. 1, 2

[15] Han Liang, Jiacheng Bao, Ruichi Zhang, Sihan Ren, Yuecheng Xu, Sibei Yang, Xin Chen, Jingyi Yu, and Lan Xu. Omg: Towards open-vocabulary motion generation via mixture of controllers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–493, 2024. 7

[16] Xiaoyang Liu, Yunyao Mao, Wengang Zhou, and Houqiang Li. Motionrl: Align text-to-motion generation to human preferences with multi-reward reinforcement learning. *arXiv preprint arXiv:2410.06513*, 2024. 2

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[18] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 5

[19] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *International Conference on Computer Vision (ICCV)*, 2023. 4, 5, 7

[20] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. PMID: 27992262. 1

[21] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big data*, 4(4):236–252, 2016. 6

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 4

[23] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 2

[24] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4, 6

[25] Xiaofeng Tan, Hongsong Wang, Xin Geng, and Pan Zhou. Sopo: Text-to-motion generation using semi-online preference optimization. *arXiv preprint arXiv:2405.14734*, 2024. 2

[26] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion

generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2

[27] Guy Tevet, Sigal Raab, Brian Gordon, Yoni Shafir, Daniel Cohen-or, and Amit Haim Bermano. Human motion diffusion model. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 2, 3, 4, 7, 8

[28] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024. 2

[29] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*, 2022. 8

[30] Weihao Yuan, Yisheng He, Weichao Shen, Yuan Dong, Xiaodong Gu, Zilong Dong, Liefeng Bo, and Qixing Huang. Mogents: Motion generation based on spatial-temporal joint modeling. *Advances in Neural Information Processing Systems*, 37:130739–130763, 2025. 2

[31] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14730–14740, 2023. 7

[32] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. Remodiffuse: Retrieval-augmented motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 364–373, 2023. 2

[33] Mingyuan Zhang, Huirong Li, Zhongang Cai, Jiawei Ren, Lei Yang, and Ziwei Liu. Finemogen: Fine-grained spatio-temporal motion generation and editing. *Advances in Neural Information Processing Systems*, 36:13981–13992, 2023. 2

[34] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):4115–4128, 2024. 2, 7, 8

[35] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, and Ziwei Liu. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, page 397–421. Springer, 2024. 7

[36] Mingyuan Zhang, Daisheng Jin, Chenyang Gu, Fangzhou Hong, Zhongang Cai, Jingfang Huang, Chongzhi Zhang, Xinying Guo, Lei Yang, Ying He, et al. Large motion model for unified multi-modal motion generation. In *European Conference on Computer Vision*, pages 397–421. Springer, 2025. 2

[37] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer Nature Switzerland, 2024. 7

[38] Zeyu Zhang, Akide Liu, Ian Reid, Richard Hartley, Bohan Zhuang, and Hao Tang. Motion mamba: Efficient and long sequence motion generation. In *European Conference on Computer Vision*, pages 265–282. Springer, 2024. 2

[39] Chongyang Zhong, Lei Hu, Zihao Zhang, and Shihong Xia. Attt2m: Text-driven human motion generation with multi-perspective attention mechanism. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 509–519, 2023. 2

[40] Qiran Zou, Shangyuan Yuan, Shian Du, Yu Wang, Chang Liu, Yi Xu, Jie Chen, and Xiangyang Ji. Parco: Part-coordinating text-to-motion synthesis. In *Computer Vision – ECCV 2024*, pages 126–143, Cham, 2025. Springer Nature Switzerland. 7

[41] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: An autoregressive multilingual sign language generator. *arXiv preprint arXiv:2411.17799*, 2024. 1, 2