SoPo: Text-to-Motion Generation Using Semi-Online Preference Optimization

Anonymous ICCV submission

Paper ID 7163

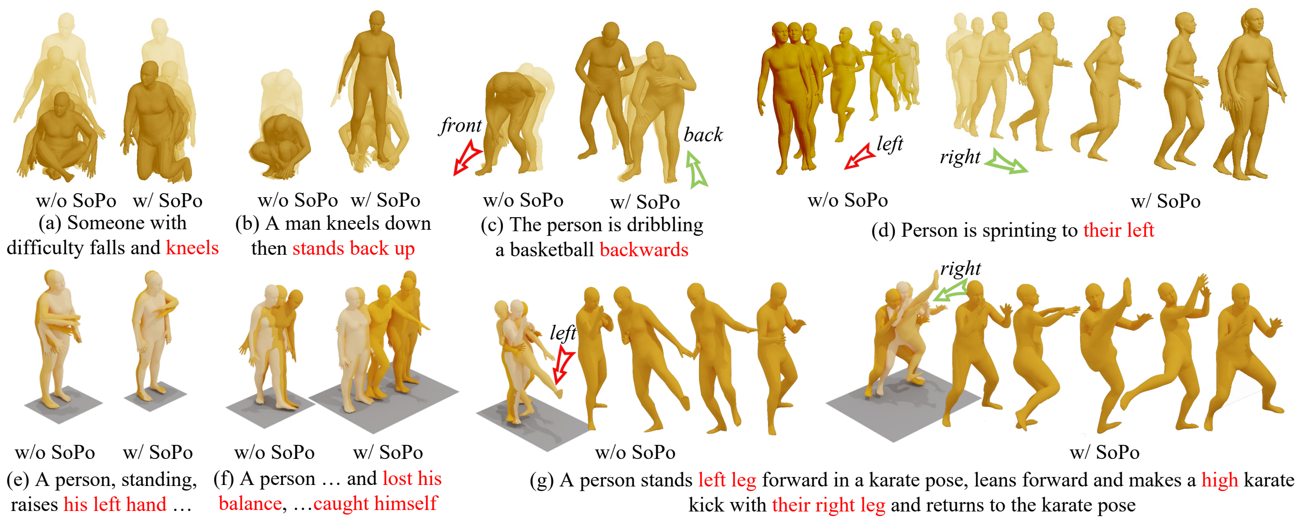

Figure 1. Visual results on HumanML3D dataset. We integrate our SoPo into MDM [28] and MLD [1], respectively. Our SoPo improves the alignment between text and motion preferences. Here, the red text denotes descriptions inconsistent with the generated motion.

Abstract

Text-to-motion generation is essential for advancing the creative industry but often presents challenges in producing consistent, realistic motions. To address this, we focus on fine-tuning text-to-motion models to consistently favor high-quality, human-preferred motions—a critical yet largely unexplored problem. In this work, we theoretically investigate the DPO under both online and offline settings, and reveal their respective limitation: overfitting in offline DPO, and biased sampling in online DPO. Building on our theoretical insights, we introduce Semi-online Preference Optimization (SoPo), a DPO-based method for training text-to-motion models using "semi-online" data pair, consisting of unpreferred motion from online distribution and preferred motion in offline datasets. This method leverages both online and offline DPO, allowing each to compensate for the other's limitations. Extensive experiments demonstrate that SoPo outperforms other preference alignment methods, with an MM-Dist of 3.25% (vs e.g. 0.76% of MoDiPO) on the MLD model, 2.91% (vs e.g. 0.66% of MoDiPO) on MDM model, respectively. Additionally, the MLD model fine-tuned by our SoPo surpasses the SoTA model in terms of R-precision and MM Dist. Visualization results also show the efficacy of our SoPo in preference alignment. The source code will be released publicly.

1. Introduction

Text-to-motion generation aims to synthesize realistic 3D human motions based on textual descriptions, unlocking numerous applications in gaming, filmmaking, virtual and augmented reality, and robotics [1, 3, 5, 8]. Recent advances in generative models [30, 31, 40], particularly diffusion models [1, 3, 10, 17, 20, 26–28, 38], have significantly improved text-to-video generation. However, text-to-motion models often encounter challenges in generating consistent, realistic motions due to several key factors.

Firstly, models are often trained on diverse text-motion pairs where descriptions vary widely in style, detail, and purpose. This variance can cause inconsistencies, producing motions that do not always meet realism or accuracy standards [24, 43]. Secondly, text-to-motion models are probabilistic, allowing diverse outputs for each description. While this promotes variety, it also increases the chances of generating undesirable variations [8]. Lastly, the complexity of coordinating multiple flexible human joints results in unpredictable outcomes, increasing the difficulty of achieving smooth and realistic motion [43]. Together, these factors limit the quality and reliability of current methods of text-to-motion generation.

In this work, we focus on refining text-to-motion models to consistently generate high-quality and human-preferred motions, a largely unexplored but essential area given its wide applicability. To our knowledge, MoDiPO [17] is the only work directly addressing this. MoDiPO applies a preference alignment method, DPO [25], originally developed for language and text-to-image models, to the text-to-motion domain. This approach fine-tunes models on datasets where each description pairs with both preferred and unpreferred motions, guiding the model toward more desirable outputs. Despite MoDiPO's promising results, challenges remain, as undesired motions continue to arise, as shown in Fig. 1. Unfortunately, this issue is still underexplored, with limited efforts directed at advancing preference alignment approaches to mitigate it effectively.

Contributions. Building upon MoDiPO, this work addresses the above problem, and derives some new results and alternatives for text-to-motion generation alignment. Particularly, we theoretically investigate the limitations of online and offline DPO, and then propose a Semi-Online Preference Optimization (SoPo) to solve the alignment issues in online and offline DPO for text-to-motion generation. Our contributions are highlighted below.

Our first contribution is the explicit revelation of the limitations of both online and offline DPO. Online DPO is constrained by biased sampling, resulting in high-preference scores that limit the preference gap between preferred and unpreferred motions. Meanwhile, offline DPO suffers from overfitting due to limited labeled preference data, especially for unpreferred motions, leading to poor generalization. This results in MoDiPO's limited and sometimes inconsistent preference alignment performance.

Inspired by our theory, we propose a novel and effective SoPo method to address these limitations. SoPo trains models on "semi-online" data pairs that incorporate high-quality preferred motions from offline datasets alongside diverse unpreferred motions generated dynamically. This blend leverages the offline dataset's human-labeled quality to counter online DPO's preference gap issues, while the dynamically generated unpreferred motions mitigate offline DPO's overfitting by increasing motion diversity.

Finally, extensive experimental results show that our SoPo significantly outperforms the SoTA baselines. For example, on the HumanML3D dataset, integrating our SoPo into MLD brings 0.222 in Diversity and 3.25% in MM Dist improvement. By comparison, combining MLD with MoDiPO only bring 0.091 and -0.01% respectively. These results underscore SoPo's effectiveness in improving human-preference alignment in text-to-motion generation.

2. Related Works

2.1. Text-to-Motion Generation

Text-to-motion generation [4, 11, 14, 20, 23, 36, 37, 41] is an important research area with significant application of computer vision. Recently, diffusion-based models have achieved substantial advances in motion generation, for enhancing both the quality and diversity of generated motion with stable training [3, 26–28]. Specifically, MotionDiffuse [38] is a pioneering text-driven diffusion model enabling fine-grained body control and flexible, arbitrary-length motion synthesis. Tevet et. al. [28] represents a transformer-based diffusion model leveraging geometric losses to improve training efficiency and performance. Chen et. al. [1] combine latent space and conditional diffusion techniques to generate human motions efficiently. Kong et. al. [10] prioritizes important motions using a discrete representation and adaptive noise schedule, reaching richer and more diverse motion generation. Dai et. al. [3] present a real-time controllable motion generation model that uses latent consistency distillation to balance efficiency and high-quality output. Although these models can generate diverse motions, they still encounter challenges in generating realistic motions that align with textual descriptions.

2.2. Direct Preference Optimization

Preference alignment seeks to learn the distribution of preferences across different outputs given the same conditions. It has achieved significant success with large-scale language models (LLMs) [7, 25], text-to-3D generation [33], and image generation [13, 15, 29, 32, 34], demonstrating promise for overcoming the aforementioned issue. These methods can generally be categorized into offline DPO [16, 29] and online DPO methods [13, 15, 32, 34]. Offline DPO methods are trained on pre-prepared offline datasets, where preference data are labeled by human annotators [29] or through AI-generated feedback [17]. In contrast, online DPO methods dynamically generate data using either a policy [13] or a reference model [34], subsequently constructing paired preference data based on feedback from humans [32] or AI [16]. Despite achieving satisfactory results in text-to-image generation, research on DPO for text-to-motion generation, akin to MoDiPO [17], remains limited. However, MoDiPO suffers from overfitting and inadequate gaps between preference data.

3. Motivation: Rethink Offline & Online DPO

Here we analyze DPO in MoDiPO to explain its inferior alignment performance for text-to-motion generation. To this end, we first briefly introduce DPO [25]. Let $\mathcal{D}$ be a preference dataset which comprises numerous triples, each containing a text condition c and a motion pair $x^w \succ x^l$ where x^w and x^l respectively denote the preferred motion and unpreferred one. With this dataset, Reinforcement Learning from Human Feedback (RLHF) [2] first trains a reward model $r(x,c)$ to access the quality of x under the condition c. Then RLHF maximizes cumulative rewards while maintaining a KL constraint between the policy model π_θ and a reference model π_{ref}:

$$\max_{\pi_\theta} \mathop{\mathbb{E}}_{c\sim\mathcal{D}, x\sim\pi_\theta(\cdot|c)} \left[r(x,c) - \beta D_{\text{KL}}\left(\pi_\theta(x|c) \,\|\, \pi_{\text{ref}}(x|c)\right)\right]. \tag{1}$$

Here one often uses the frozen pretrained model as the reference model π_{ref} and current trainable text-to-motion model as the policy model π_θ.

Building upon RLHF, DPO [25] analyzes the close solution of problem in (1) to simplify RLHF's loss as follows:

$$\mathcal{L}_{\text{DPO}}(\theta) = \mathbb{E}_{(x^w,x^l,c)\sim\mathcal{D}}\left[-\log\sigma\left(\beta\mathcal{H}_\theta(x^w,x^l,c)\right)\right], \tag{2}$$

where $\mathcal{H}_\theta(x^w,x^l,c) = h_\theta(x^w,c) - h_\theta(x^l,c)$, $h_\theta(x,c) = \log\frac{\pi_\theta(x|c)}{\pi_{\text{ref}}(x|c)}$, and σ is the logistic function. When there are multiple preferred motions (responses) under a condition c, i.e., $x^1 \succ x^2 \succ \cdots \succ x^K$ $(K \geq 2)$, by using Plackett-Luce model [21], DPO can be extended as:

$$\mathcal{L}_{\text{off}}(\theta) = -\mathbb{E}_{(x^{1:K},c)\sim\mathcal{D}}\left[\log\prod_{k=1}^{K}\frac{\exp(\beta h_\theta(x^k,c))}{\sum_{j=k}^{K}\exp(\beta h_\theta(x^j,c))}\right]. \tag{3}$$

When $K=2$, $\mathcal{L}_{\text{off}}$ degenerates to $\mathcal{L}_{\text{DPO}}$. Since MoDiPO uses multiple preferred motions for alignment, we will focus on analyze the general formulation in Eq. (3).

3.1. Offline DPO

In Eq. (3), its training samples are sampled from an offline dataset $\mathcal{D}$. Accordingly, DPO in Eq. (3) is also called "offline DPO". Here we analyze its preference optimization.

Theorem 1. *Given a preference motion dataset $\mathcal{D}$, a reference model π_{ref}, and ground-truth preference distribution p_{gt}, the gradient of $\nabla_\theta\mathcal{L}_{\text{off}}$ can be written as:*

$$\nabla_\theta\mathcal{L}_{\text{off}}(\theta) = \mathbb{E}_{(x^{1:K},c\sim\mathcal{D})}\nabla_\theta D_{KL}(p_{\text{gt}}||p_\theta). \tag{4}$$

Here $p_\theta(x^{1:K}|c) = \prod_{k=1}^{K} p_\theta(x^k|c)$ with represents the likelihood that policy model generates motions $x^{1:K}$matching their rankings, where $p_\theta(x^k|c) = \frac{(\exp h_\theta(x^k,c))^\beta}{\sum_{j=k}^{K}(\exp h_\theta(x^j,c))^\beta}$.

See its proof in Appendix B.1.Theorem 1 shows that the gradient of offline DPO aligns with the gradient of the forward KL divergence, $D_{KL}(p_{\text{gt}}||p_\theta)$. This suggests that the policy model p_θ (i.e., the trainable text-to-motion model) is optimized to match its text-to-motion distribution with the ground-truth motion preference distribution p_{gt}.

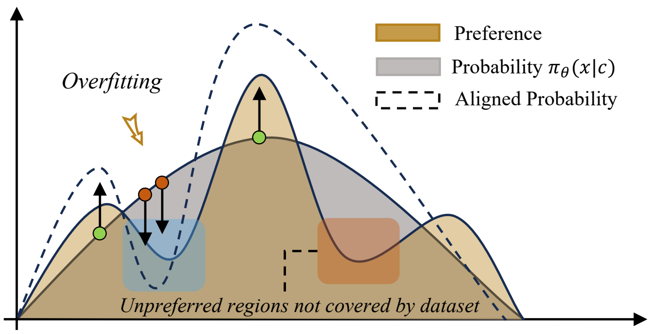

Figure 2. Illustration of overfitting issue in offline DPO. The blue and red points represent preferred and unpreferred motion, respectively. The blue region represents the overfitting distribution that arises from repeatedly optimizing with the same unpreferred motions in each iteration. The red region denotes the neglected unpreferred area uncovered by the offline dataset.

However, because training data comes from a fixed dataset $\mathcal{D}$, the model risks overfitting on this static set, especially for unpreferred data. Due to limited labeled data, text-to-motion datasets usually include only one preferred motion group $x_c^{1:K}$ per condition c, making $p_{\text{gt}}(\cdot|c)$ approximate a one-point distribution i.e. $p_{\text{gt}}(x_c^{1:K}|c) = 1$. In this case, minimizing $D_{KL}(p_{\text{gt}}||p_\theta)$ is equivalent to maximizing the likelihood p_θ via $\min D_{KL}(p_{\text{gt}}||p_\theta) \Leftrightarrow \min -\log p_\theta(x^{1:K}|c)$. This drives offline DPO to progressively increase the probability $p_\theta(x_c^{1:K}|c)$, iteratively widening the preference gap between preferred and unpreferred motions. As shown in Fig. 2, throughout iterations, the model predominantly learns from the single fixed motion group $x_c^{1:K}$ associated with a given c, causing the preference gap within $x_c^{1:K}$ to expand. This issue, observed in [42] as well, implies that when unpreferred data is scarce, offline DPO enables the model to avoid only specific unpreferred patterns (e.g., the blue regions in Fig. 2) but disregards frequently occurring unpreferred motions (e.g., the red regions in Fig. 2).

Despite this limitation, the offline dataset is manually labeled and provides valuable preference information, where the gap between preferred and unpreferred motions is large, benefiting learning preferred motions.

3.2. Online DPO

Per training iteration in online DPO, one uses current policy model π_θ to generate K samples for a text c, and adopts a well-trained reward model r to rank them in terms of their preference so that $x_{\pi_\theta}^1 \succ x_{\pi_\theta}^2 \succ \cdots \succ x_{\pi_\theta}^K$, where $x_{\pi_\theta}^i$ is a sample generated by π_θ without gradient propagation to parameter θ. Then with Plackett-Luce model [21], the probability of $x_{\pi_\theta}^k$ being the k-th preferred sample is as:

$$p_r(x_{\pi_\theta}^k|c) = \frac{\exp r(x_{\pi_\theta}^k, c)}{\sum_{i=k}^{K}\exp r(x_{\pi_\theta}^i, c)}. \tag{5}$$

Then we can analyze online DPO below.

Theorem 2. *Given a reward model r and a reference model π_{ref}, then for the online DPO loss $\mathcal{L}_{\text{on}}$, its gradient is:*

$$\nabla_\theta \mathcal{L}_{\text{on}}(\theta) = \mathbb{E}_{c\sim\mathcal{D}} \nabla_\theta\, p_{\bar{\pi}_\theta}(x^{1:K}|c) D_{KL}(p_r||p_\theta), \quad (6)$$

where $p_{\bar{\pi}_\theta}(x^{1:K}|c) = \prod_{k=1}^{K} p_{\bar{\pi}_\theta}(x^k|c)$ with $p_{\bar{\pi}_\theta}(x^k|c)$ being the generative probability of policy model to generate x^k conditioned on c, and $p_\theta(x^k) = \frac{(\exp h_\theta(x_k,c))^\beta}{\sum_{j=k}^{K}(\exp h_\theta(x_j,c))^\beta}$ denotes the likehood that policy model generates motion x_k with the k-th largest probability.

See the proof in Appendix B.2 Theorem 2 indicates that online DPO minimizes the forward KL divergence $D_{KL}(p_r||p_\theta)$. Thus, online DPO trains the policy model π_θ, i.e., the text-to-motion model, to align its text-to-motion distribution with the online preference distribution $p_r(x|c)$.

We now discuss the training bias and limitations. Specifically, motions with higher generative probability, $p_{\bar{\pi}_\theta}(x_{\bar{\pi}_\theta}|c)$, are commonly synthesized and thus frequently used to train the model π_θ. In contrast, motions with lower generative probability are seldom generated and rarely contribute to training, despite potentially high human preference. Indeed, when the generative probability of a sample $x_{\bar{\pi}_\theta}$ is low but its reward $r(x_{\bar{\pi}_\theta}, c)$ is high, the gradient still approaches zero: $\lim_{p_{\pi_\theta}(x_{\bar{\pi}_\theta}|c)\to 0, r(x_{\bar{\pi}_\theta},c)\to 1} \nabla_\theta \mathcal{L}_{\text{on}} = \mathbf{0}$ (see derivation in Appendix B.2). This implies that even highly valuable preferred motions are overlooked by online DPO which primarily trains the text-to-motion model on frequent samples, regardless of preference.

Additionally, online DPO aligns generative probability $p_{\bar{\pi}_\theta}(x_{\bar{\pi}_\theta}|c)$ with preference distribution $p_r(x_{\bar{\pi}_\theta}|c)$, resulting in their positive correlation. So motions with relatively high generative probabilities also tend to exhibit high preferences. However, these high-preference motions are ranked by a reward model, which causes half of them—those with lower rankings k yet relatively high preference scores $r(x^k_{\bar{\pi}_\theta}, c)$—to be treated as unpreferred. As a result, many unpreferred training motions still exhibit substantial preferences, narrowing the gap between preferred and unpreferred motions compared to manually labeled offline datasets.

On the other hand, online DPO continuously generates diverse data, especially unpreferred samples, in each iteration, providing varied preference information for the text-to-motion model. This dynamic data generation mitigates overfitting mentioned in offline DPO, and encourages the model to learn to avoid multiple types of undesired patterns.

3.3. DPO in MoDiPO for Text-to-Motion

DPO in MoDiPO [17] uses an offline dataset $\mathcal{D}$ which is indeed generated by a pre-trained model π_p, denoted as:

$$\begin{aligned} x^w_{\pi_p} &= \text{argmax}_{x^{1:K}_{\pi_p}\in\bar{\pi}_p} \exp r(x^k_{\pi_p}, c), \\ x^l_{\pi_p} &= \text{argmin}_{x^{1:K}_{\pi_p}\in\bar{\pi}_p} \exp r(x^k_{\pi_p}, c), \quad (7) \\ \mathcal{D} &= \{(x^w_{\pi_p}, x^l_{\pi_p}, c)|c \in \text{offline textural sets}\}. \end{aligned}$$

For discussion, we formulate its sampled distribution as:

$$p^{Mo}_{\text{gt}}(x_w, x_l|c) = \mathbb{I}((x_w, x_l, c) \in \mathcal{D}), \quad (8)$$

where the indication function $\mathbb{I}(\mathcal{E}) = 1$ if event $\mathcal{E}$ happens; otherwise, $\mathbb{I}(\mathcal{E}) = 0$.

From Eq. (7), we observe that, like online DPO, MoDiPO samples preference motions from the distribution $p_{\pi_p}(x|c)$ induced by the pre-trained model π_p. This leads to two main issues like online DPO. 1) Samples with low generative probability $p_{\pi_p}(x|c)$ but high preferences $r(x,c)$ are rarely generated by π_p and thus seldom contribute to training, even though they are highly desirable motions. 2) As discussed in Sec. 3.2, the motions x_{π_p} generated by π_p typically exhibit both high generative probability and preference scores, which causes half of the preferred samples to be selected as unpreferred, skewing the model's learning process. See the detailed discussion in Sec. 3.2.

Additionally, from Eq. (8), we see that for a given condition c, MoDiPO trains on fixed preference data, similar to offline DPO. Consequently, MoDiPO is limited to avoiding only the unpreferred motions valued by the pre-trained model π_p, rather than those relevant to the policy model π_θ. Thus, MoDiPO inherits the limitations of both online and offline DPO, which constrains its alignment performance.

4. Semi-Online Preference Optimization

4.1. Overview of SoPo

We introduce our Semi-Online Preference Optimization (SoPo) to address the limitations in both online and offline DPO for text-to-motion generation. Its core idea is to train the text-to-motion model on semi-online data pairs, where high-preference motions are from offline datasets, while low-preference and high-diversity unpreferred motions are generated online.

As analyzed in Sec. 3, offline DPO provides high-preference motions with a clear preference gap from unpreferred ones, but suffers from overfitting due to reliance on fixed single-source unpreferred motions. Online DPO, in contrast, benefits from diverse and dynamically generated data but often lacks a sufficient preference gap between preferred and unpreferred motions, and overlooks low-probability preferred motions. To capitalize on the strengths of both, SoPo samples diverse unpreferred motions $x^l_{\bar{\pi}_\theta}$ from online generation and high-preference motions $x^w_{\mathcal{D}}$ from offline datasets, ensuring a broad preference gap between preferred and unpreferred motions. Consequently, SoPo overcomes the overfitting issues of offline

DPO and the insufficient preference gaps of online DPO. Accordingly, we can arrive at our SoPo:

$$\mathcal{L}_{\mathrm{DSoPo}}(\theta) = -\,\mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^l\sim\bar{\pi}_\theta(x|c)}\log\sigma\Big(\beta\mathcal{H}_\theta(x^w,x^l,c)\Big), \tag{9}$$

where $\mathcal{H}_\theta(x^w, x^l, c)$ is defined below Eq. (2), x^w is preferred motion from the offline dataset, and x^l is unpreferred motion sampled from online DPO.

However, direct online generation of unpreferred motions from the policy model presents challenges, given the positive correlation between the generative distribution $p_{\bar{\pi}_\theta}$ and preference distribution p_r. Additionally, a large gap between preferred and unpreferred motions remains essential for effective SoPo. In Sec. 4.2 and 4.3, we receptively elaborate on SoPo's designs to address these challenges.

4.2. Online Generation for Unpreferred Motions

Here we introduce our generation pipeline for diverse unpreferred motions. Specifically, given a condition c, we first generate K motions $\{x^k_{\bar{\pi}_\theta}\}_{k=1}^K$ from the policy model π_θ, and select the one with the lowest preference value:

$$x^l_{\bar{\pi}_\theta} = \mathrm{argmin}_{\{x^k_{\bar{\pi}_\theta}\}_{k=1}^K\sim\pi_\theta} r(x^k_{\bar{\pi}_\theta}, c). \tag{10}$$

However, $x^l_{\bar{\pi}_\theta}$ could still exhibit a relatively high preference $r(x^l_{\bar{\pi}_\theta}, c)$ due to the positive correlation between the generative probability $p_{\bar{\pi}_\theta}$ and preference distribution p_r (see Sec. 3.2 or 3.3). To ensure the selection of genuinely unpreferred motions, we apply a threshold τ and check if any motion in $\{x^k_{\bar{\pi}_\theta}\}_{k=1}^K$ has a preference score below τ. It results in two possible cases with respective training strategies:

Case 1: The group $\{x^k_{\bar{\pi}_\theta}\}_{k=1}^K$ contains a low-preference unpreferred motion $x^l_{\bar{\pi}_\theta}$. Then we select these unpreferred motions iteratively which ensure diversity due to randomness of online generations and address the diversity lacking issue in offline DPO.

Case 2: The group does not has low-preference unpreferred motion $x^l_{\bar{\pi}_\theta}$. So all sampled motions in this group have high preference, and should be classified as high-preference rather than unpreferred. It indicates that the generative model performs well on condition c without notable unpreferred flaws. So training should focus on high-quality preferred motions from offline datasets to further improve generative quality.

To operationalize this, we apply a two-step process: 1) distribution separation and 2) training loss amendment.

1) Distribution separation: With a threshold τ, we separate the distribution $p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)$ into two sub-distributions:

$$p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c) = \underbrace{p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)p_\tau(r(x^l_{\bar{\pi}_\theta},c)\geq\tau)}_{\text{relatively high-preference unpreferred motions } \bar{\pi}^{hu}_\theta} + \underbrace{p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)p_\tau(r(x^l_{\bar{\pi}_\theta},c)<\tau)}_{\text{valuable unpreferred motions } \bar{\pi}^{vu}_\theta}, \tag{11}$$

where $p_{\bar{\pi}_\theta}(x^{1:K}|c) = \prod_{k=1}^K p_{\bar{\pi}_\theta}(x^k|c)$, $p_{\bar{\pi}_\theta}(x^k|c)$ is the generative probability of policy model π_θ to generate x^k conditioned on c, $p_\tau(r(x^l_{\bar{\pi}_\theta},c)\geq\tau)$ is the probability of the event $x^l_{\bar{\pi}_\theta}\geq\tau$, and $p_\tau(r(x^l_{\bar{\pi}_\theta},c)\leq\tau)$ has similar meaning.

Eq. (11) indicates that the online generative distribution $\bar{\pi}_\theta(x^{1:K}_{\bar{\pi}_\theta}|c)$ can be separated according to whether the sampled motion $x^{1:K}_{\bar{\pi}_\theta}$ group contains valuable unpreferred motions. Accordingly, our objective loss in Eq. (9) can also be divided into two ones: $\mathcal{L}_{\mathrm{DSoPo}}(\theta) = \mathcal{L}_{\mathrm{vu}}(\theta) + \mathcal{L}_{\mathrm{hu}}(\theta)$, where $\mathcal{L}_{\mathrm{vu}}(\theta)$ targets valuable unpreferred motions and $\mathcal{L}_{\mathrm{hu}}(\theta)$ targets high-preference unpreferred motions:

$$\begin{aligned}\mathcal{L}_{\mathrm{vu}} &= -\,\mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{vu}(c)\mathbb{E}_{x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}^{vu*}_\theta(\cdot|c)}\log\sigma\big(\beta\mathcal{H}_\theta(x^w,x^l_{\bar{\pi}_\theta},c)\big),\\ \mathcal{L}_{\mathrm{hu}} &= -\,\mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{hu}(c)\mathbb{E}_{x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}^{hu*}_\theta(\cdot|c)}\log\sigma\big(\beta\mathcal{H}_\theta(x^w,x^l_{\bar{\pi}_\theta},c)\big),\end{aligned} \tag{12}$$

where $\mathcal{H}_\theta(x^w, x^l_{\bar{\pi}_\theta}, c)$ is defined in Eq. (2), $p_{\bar{\pi}^{vu*}_\theta}(\cdot) = \frac{p^{vu}_{\bar{\pi}_\theta}(\cdot)}{Z_{vu}(c)}$ and $p^{hu*}_{\bar{\pi}_\theta}(\cdot) = \frac{p^{hu}_{\bar{\pi}_\theta}(\cdot)}{Z_{hu}(c)}$ respectively denote the distributions of valuable unpreferred and high-preference unpreferred motions. Here $Z_{vu}(c) = \int p_{\bar{\pi}^{vu}_\theta}(x)\mathrm{d}x$ and $Z_{hu}(c) = \int p_{\bar{\pi}^{hu}_\theta}(x)\mathrm{d}x$ are the partition functions, and are unnecessary to be computed in our implementation (Appendix B.3).

In Eq. (12), $\mathcal{L}_{\mathrm{vu}}(\theta)$ and $\mathcal{L}_{\mathrm{hu}}(\theta)$ respectively denote the loss on valuable unpreferred and relatively high-preference unpreferred motions. If the probability $p_{\bar{\pi}^{vu}_\theta}(x)$ that the sampled motion group contains valuable unpreferred motion is high, $Z_{vu}(c)$ will also be large. In this case, $\mathcal{L}_{\mathrm{DSoPo}}(\theta)$ will pay more attention to optimizing $\mathcal{L}_{\mathrm{vu}}(\theta)$ to learn not generating these unpreferred motions.

2) Training loss amendment: As discussed above, unpreferred motions in case 2 have relatively high-preference (score $\geq r$), and thus should be classified into preferred motions for training. Accordingly, we rewrite the loss $\mathcal{L}_{\mathrm{hu}}(\theta)$ into $\mathcal{L}_{\mathrm{USoPo}}(\theta)$ for training on preferred motions:

$$\begin{aligned}\mathcal{L}_{\mathrm{USoPo-hu}}(\theta) &= -\mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{hu}(c)\log\sigma\Big(\beta h_\theta(x^w,c)\Big),\\ \mathcal{L}_{\mathrm{USoPo}}(\theta) &= \mathcal{L}_{\mathrm{USoPo-hu}}(\theta) + \mathcal{L}_{\mathrm{vu}}(\theta).\end{aligned} \tag{13}$$

See more discussion on $\mathcal{L}_{\mathrm{USoPo}}/\mathcal{L}_{\mathrm{DSoPo}}$ in Appendix B.4.

4.3. Offline Sampling for Preferred Motions

As aforementioned, online DPO is plagued by an insufficiency preference gap between unpreferred and preferred motions. Fortunately, high-quality preferred motions from offline datasets can help alleviate this issue. However, these preferred motions do not always exhibit a significant gap with the generated motions, especially when the model and offline datasets are well aligned. Thus, preferred motions with a larger preference gap from generated unpreferred motions, as identified in Sec. 4.2, are particularly valuable and should be emphasized during training.

To utilize the generated unpreferred motion set $\mathcal{D}_c$ conditioned on c from Sec. 4.2, we calculate its proximity with the unpreferred motions in $\mathcal{D}_c$ using cosine similarity:

$$S(x^w) = \min_{x^k_{\bar{\pi}_\theta} \sim \mathcal{D}_c} \cos(x^w, x^k_{\bar{\pi}_\theta}).$$

Then we reweight the loss of the preferred motions by using $\beta_w(x_w) = \beta(C - S(x^w))$ with a constant $C \geq 1$:

$$\begin{aligned}\mathcal{L}_{\mathrm{SoPo}}(\theta) = &- \mathbb{E}_{(x^w,c)\sim\mathcal{D}, x^{1:K}_{\bar{\pi}_\theta} \sim \bar{\pi}^{vu*}_\theta(\cdot|c)} Z_{vu}(c) \\ &\Big[\log\sigma\Big(\beta_w(x^w) h_\theta(x^w, c) - \beta h_\theta(x^l, c)\Big)\Big] \\ &- \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \log\sigma\Big(\beta_w(x^w) h_\theta(x^w, c)\Big).\end{aligned} \tag{14}$$

As similar samples have similar preferences, this reweighting strategy guides the model to prioritize preferred motions with a significant preference gap from unpreferred ones. Accordingly, this reweighting strategy relieves and even addresses the small preference gap issue in online DPO.

4.4. SoPo for Diffusion-Based Text-to-Motion

Recently, diffusion text-to-motion models have achieved remarkable success [3, 26, 27, 31], enabling the generation of diverse and realistic motion sequences. Inspired by [29], we derive the objective function of SoPo for diffusion-based text-to-image generation (See proof in Appendix B.5):

$$\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo}} = \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}} + \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}}, \tag{15}$$

$$\begin{aligned}\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}} = &-\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}, x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}^{vu*}_\theta(\cdot|c)} Z_{vu}(c) \\ &\Big[\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\mathrm{ref},x^w_t) - \beta\mathcal{L}(\theta,\mathrm{ref},x^l_t))\Big)\Big] \\ \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}} = &-\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \\ &\Big[\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\mathrm{ref},x^w_t)\Big)\Big]\end{aligned} \tag{16}$$

where $\mathcal{L}(\theta, \mathrm{ref}, x_t) = \mathcal{L}(\theta, x_t) - \mathcal{L}(\mathrm{ref}, x_t)$, and $\mathcal{L}(\theta/\mathrm{ref}, x_t) = \|\epsilon_{\theta/\mathrm{ref}}(x_t, t) - \epsilon\|^2_2$ denotes the loss of the policy or reference model. Equivalently, we optimize the following form

$$\begin{aligned}\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo}}(\theta) = &-\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}, x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}_\theta(\cdot|c)} \\ &\begin{cases}\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)\mathcal{L}(\theta,\mathrm{ref},x^w_t) \\ \quad -\beta\mathcal{L}(\theta,\mathrm{ref},x^l_t))\Big), & \text{if } r(x^l,c) < \tau, \\ \log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\mathrm{ref},x^w_t)\Big), & \text{otherwise.}\end{cases}\end{aligned} \tag{17}$$

where $x^l = \mathrm{argmin}_{\{x^k_{\bar{\pi}_\theta}\}^K_{k=1}\sim\pi_\theta} r(x^k_{\pi_\theta}, c)$. See proof and more details like the pipeline in Appendix A.

5. Experiment

Datasets We evaluate our SoPo on two widely used benchmarks for text-driven motion generation, the HumanML3D [5] and the KIT-ML [22] dataset.

Evaluation metrics We evaluate our results across two key dimensions: alignment quality and generation quality. Consistent with prior work [3, 26, 40], alignment quality is measured using R-Precision and MM Dist, while generation quality is assessed via Diversity and FID.

Implementation details Given the scarcity of preference data in the motion domain, we use the training set of humanML3D as the offline preferred motion dataset. To online generate unpreferred motions, we employ TMR, a text-to-motion retrieval model [19], as the reward model. Values of K and τ were tuned based on preliminary experiments to balance performance and computational cost. We set the cut-off threshold to $\tau = 0.45$, the hyper-parameters in Eq. to (14) $C = 2$ and $\beta = 1$, and the number of online sampled motions to $K = 4$ for MDM [18], $K = 2$ for MLD [1]. Each model was trained for no more than 100 minutes on a single NVIDIA GeForce RTX 4090D GPU. Note that MLD* [3] is specifically designed for HumanML3D and not adapted for KIT-ML, where we use the original MLD [1] instead. More training details can be found in Appendix C.1.

5.1. Main Results

Settings To evaluate our SoPo for preference alignment and motion generation, we compare it with SoTA methods, including preference alignment methods [17], and text-to-motion generation methods [1, 9, 10, 18, 26–28, 30, 31, 38, 40]. To ensure a fair comparison with preference alignment methods, we utilize our SoPo to fine-tune MLD [1] and MDM [28]. Since motion sampling is included in our pipeline, we adopt a fast variant [28] with only 50 sampling steps to reduce computational costs. To further demonstrate the effectiveness of our SoPo, we fine-tune MLD*, a performance-enhanced version reproduced by [3], to compare with SoTA text-to-motion methods. Since MLD* [3] is not adapted for the KIT-ML dataset, we use the original MLD [1] and MoMask [6] to represent diffusion-based and autoregressive models, respectively, for this dataset.

Comparison with preference alignment methods Table 1 presents the results of preference alignment methods. MoDiPO is a DPO-based method for motion generation. Due to the limitation of offline and online DPO [25], it suffers from overfitting and biased sampling in aligning AI preferences. Unlike classical DPO methods, our SoPo aims to effectively utilize diverse high-probability unpreferred motions and high-quality preferred motions for training, thereby improving generation quality and reducing the likelihood of generating common unpreferred motions.

Hence, our SoPo achieves the best performance across all metrics except for FID, particularly excelling in preference alignment metrics such as R-Precision and MM Dist. Specifically, our SoPo achieves an improvement of up to 5.27%, 4.50%, and 2.80% in R-Precision, significantly surpassing that of the suboptimal method (improvement of 0.42%). It also yields a robust MM Dist enhancement of 3.25%, significantly outperforming MoDiPO's range of -12.4% to $+0.76\%$. For generation quality metrics, such as

 ICCV #7163

Methods	Time*	R-Precision ↑			MM Dist ↓	Diversity →	FID ↓
		Top 1	Top 2	Top 3			
Real	-	$.511^{\pm.003}$	$.703^{\pm.003}$	$.797^{\pm.002}$	$2.974^{\pm.008}$	$9.503^{\pm.065}$	$.002^{\pm.000}$
MLD [1]	+0 X	-	-	$.755^{\pm.003}$	$3.292^{\pm.010}$	$9.793^{\pm.072}$	$.459^{\pm.011}$
+ MoDiPO-T [17]	+121K X	-	-	$.758^{\pm.002}_{+0.40\%}$	$3.267^{\pm.010}_{+0.76\%}$	$9.747^{\pm.073}_{+0.046}$	$.303^{\pm.031}_{+33.9\%}$
+ MoDiPO-G [17]	+121K X	-	-	$.753^{\pm.003}_{-0.26\%}$	$3.294^{\pm.010}_{-0.01\%}$	$9.702^{\pm.075}_{+0.091}$	$.281^{\pm.031}_{+38.8\%}$
+ MoDiPO-O [17]	-	-	-	$.677^{\pm.003}_{-10.3\%}$	$3.701^{\pm.013}_{-12.4\%}$	$9.241^{\pm.079}_{-0.018}$	$.276^{\pm.007}_{+39.9\%}$ †
+ SoPo (Ours)	+20 X	-	-	$.763^{\pm.003}_{+1.06\%}$	$3.185^{\pm.012}_{+3.25\%}$ †	$9.525^{\pm.065}_{+0.268}$ †	$.374^{\pm.007}_{+18.5\%}$
MDM [28]	+0 X	$.418^{\pm.005}$	$.604^{\pm.005}$	$.703^{\pm.005}$	$3.658^{\pm.025}$	$9.546^{\pm.066}$	$.501^{\pm.037}$
+ MoDiPO-T [17]	+121K X	-	-	$.706^{\pm.004}_{+0.42\%}$	$3.634^{\pm.026}_{+0.66\%}$	$9.531^{\pm.073}_{+0.015}$	$.451^{\pm.031}_{+9.98\%}$
+ MoDiPO-G [17]	+121K X	-	-	$.704^{\pm.001}_{+0.14\%}$	$3.641^{\pm.025}_{+0.46\%}$	$9.495^{\pm.071}_{+0.035}$	$.486^{\pm.031}_{+2.99\%}$
MDM (fast) [28]	+0 X	$.455^{\pm.006}$	$.645^{\pm.007}$	$.749^{\pm.004}$	$3.304^{\pm.023}$	$9.948^{\pm.084}$	$.534^{\pm.052}$
+ SoPo (Ours)	+60 X	$.479^{\pm.006}_{+5.27\%}$ †	$.674^{\pm.005}_{+4.50\%}$ †	$.770^{\pm.006}_{+2.80\%}$ †	$3.208^{\pm.025}_{+2.91\%}$	$9.906^{\pm.083}_{+0.042}$	$.480^{\pm.046}_{+10.1\%}$

Table 1. **Quantitative results of preference alignment methods for text-to-motion generation on the HumanML3D test set.** Results are borrowed from those reported in [17]. Symbols "↑", "↓", and "→" indicate better performance with higher, lower, or closer-to-real-data values, respectively. Each cell is reported as $a^{\pm b}_{\pm c}$, where a is the metric value, b is the 95% confidence interval following [5], and $+c/-c$ denotes the relative performance change. Superscript "†" marks the largest improvement across all models; gray background highlights the largest improvement for each model. "Time*" denotes estimated online/offline motion generation time, with 1X" as the time for MLD [1] to generate all HumanML3D motions and "K" (unspecified in [17], typically 2∼6) as the number of motion pairs.

Methods	Year	R-Precision ↑				MM Dist ↓	Diversity →	FID ↓
		Top 1	Top 2	Top 3	Avg.			
Real	-	$0.511^{\pm0.003}$	$0.703^{\pm0.003}$	$0.797^{\pm0.002}$	0.670	$2.794^{\pm0.008}$	$9.503^{\pm0.065}$	$0.002^{\pm0.000}$
TEMOS [18]	2022	$0.424^{\pm0.002}$	$0.612^{\pm0.002}$	$0.722^{\pm0.002}$	0.586	$3.703^{\pm0.008}$	$8.973^{\pm0.071}$	$3.734^{\pm0.028}$
T2M [5]	2022	$0.457^{\pm0.002}$	$0.639^{\pm0.003}$	$0.740^{\pm0.003}$	0.612	$3.340^{\pm0.008}$	$9.188^{\pm0.002}$	$1.067^{\pm0.002}$
MDM [28]	2022	$0.418^{\pm0.005}$	$0.604^{\pm0.005}$	$0.703^{\pm0.005}$	0.575	$3.658^{\pm0.025}$	$9.546^{\pm0.066}$	$0.501^{\pm0.037}$
MLD [1]	2023	$0.481^{\pm0.003}$	$0.673^{\pm0.003}$	$0.772^{\pm0.002}$	0.642	$3.196^{\pm0.016}$	$9.724^{\pm0.082}$	$0.473^{\pm0.013}$
Fg-T2M [30]	2023	$0.418^{\pm0.005}$	$0.626^{\pm0.004}$	$0.745^{\pm0.004}$	0.596	$3.114^{\pm0.015}$	$10.930^{\pm0.083}$	$0.571^{\pm0.047}$
M2DM [10]	2023	$0.416^{\pm0.004}$	$0.628^{\pm0.004}$	$0.743^{\pm0.004}$	0.596	$3.015^{\pm0.017}$	$11.417^{\pm0.082}$	$0.515^{\pm0.029}$
MotionGPT [9]	2023	$0.492^{\pm0.003}$	$0.681^{\pm0.003}$	$0.778^{\pm0.002}$	0.650	$3.096^{\pm0.008}$	$\mathbf{9.528}^{\pm0.071}$	$0.232^{\pm0.008}$
MotionDiffuse [38]	2024	$0.491^{\pm0.004}$	$0.681^{\pm0.002}$	$0.782^{\pm0.001}$	0.651	$3.113^{\pm0.018}$	$9.410^{\pm0.049}$	$0.630^{\pm0.011}$
OMG [12]	2024	-	-	$0.784^{\pm0.002}$	-	-	$9.657^{\pm0.085}$	$0.381^{\pm0.008}$
Wang et. al. [31]	2024	$0.433^{\pm0.007}$	$0.629^{\pm0.007}$	$0.733^{\pm0.006}$	0.598	$3.430^{\pm0.061}$	$9.825^{\pm0.159}$	$0.352^{\pm0.109}$
MoDiPO-T [17]	2024	-	-	$0.758^{\pm0.002}$	-	$3.267^{\pm0.010}$	$9.747^{\pm0.073}$	$0.303^{\pm0.031}$
PriorMDM [27]	2024	0.481	-	-	-	5.610	9.620	0.600
LMM-T[1] [39]	2024	$0.496^{\pm0.002}$	$0.685^{\pm0.002}$	$0.785^{\pm0.002}$	0.655	$3.087^{\pm0.012}$	$9.176^{\pm0.074}$	$0.415^{\pm0.002}$
CrossDiff[3] [26]	2024	-	-	0.730	-	3.358	9.577	0.281
Motion Mamba [40]	2024	$0.502^{\pm0.003}$	$0.693^{\pm0.002}$	$0.792^{\pm0.002}$	0.662	$3.060^{\pm0.009}$	$9.871^{\pm0.084}$	$0.281^{\pm0.011}$
MLD* [1, 3]	2023	$0.504^{\pm0.002}$	$0.698^{\pm0.003}$	$0.796^{\pm0.002}$	0.666	$3.052^{\pm0.009}$	$9.634^{\pm0.064}$	$0.450^{\pm0.011}$
MLD* [3] + SoPo	-	$\mathbf{0.528}_{+4.76\%}$	$\mathbf{0.722}_{+3.44\%}$	$\mathbf{0.827}_{+3.89\%}$	$\mathbf{0.692}_{+3.90\%}$	$\mathbf{2.939}_{+3.70\%}$	$9.584_{+38.1\%}$	$\mathbf{0.174}_{+61.3\%}$

Table 2. **Quantitative comparison of state-of-the-art text-to-motion generation on the HumanML3D test set. Best** results are in **bold**. In the last row, subscripts "$_{+c}$" indicate performance improvements. 'MLD*' refers to the enhanced reproduction of MLD [1] from [3]. For a fair comparison, we selected the "LMM-T" [39] with a similar size to ours.

Diversity and FID, SoPo achieves a Diversity gain of 0.268, exceeding MoDiPO's -0.018 to 0.091. Although MoDiPO outperforms SoPo in FID, their results remain comparable. This can be attributed to SoPo's cautious strategy, which prioritizes learning from low-probability samples over all unpreferred motions, ensuring steady, albeit limited, improvements. Additionally, SoPo leverages original training set samples as preferred motions and generates unpreferred motions online, removing reliance on extra pairwise datasets. Moreover, SoPo's preference motion generation time is roughly 1/10 that of MoDiPO.

Comparison with motion generation methods We evaluate our SoPo method against state-of-the-art methods on the HumanML3D [5], with results reported in Table 2. Leveraging preference alignment, SoPo surpasses existing methods across multiple metrics, particularly in Top-k R-Precision, MM Dist, and FID. Notably, SoPo achieves the **best performance** in R-Precision, MM Dist, and FID. Although MotionGPT [9] exceeds SoPo by 0.056 in Diversity (9.528 vs. 9.584), our method outperforms it with a substantial R-Precision of 6.46% (0.692 v.s. 0.650), an FID gain of 33.5% (0.232 v.s. 0.174), and an MM Dist reduction

Method	Year	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
TEMOS [18]	2022	0.370	0.569	0.693	2.770	3.401	10.91
T2M [5]	2022	0.361	0.559	0.681	3.022	2.052	10.72
MLD [1]	2023	0.390	0.609	0.734	0.404	3.204	10.80
T2M-GPT [35]	2023	0.416	0.627	0.745	0.514	3.007	10.86
MotionGPT [9]	2023	0.366	0.558	0.680	0.510	3.527	10.35
MotionDiffuse[38]	2024	0.417	0.621	0.739	1.954	2.958	11.10
Mo.Mamba [40]	2024	0.419	0.645	0.765	0.307	3.021	**11.02**
MoMask [6]	2024	0.433	0.656	0.781	0.204	**2.779**	10.71
MLD [1]+ SoPo	-	0.412	0.646	0.759	0.384	3.107	10.93
MoMask [6] + SoPo	-	**0.446**	**0.673**	**0.797**	**0.176**	2.783	10.96

Table 3. **Comparison of text-to-motion generation performance on the KIT-ML dataset. Bold** highlights the best results.

Methods	R-Precision ↑			MM Dist ↓	Diversity →	FID ↓
	Top 1	Top 2	Top 3			
MDM (fast) [28]	.455	.645	.749	3.304	9.948	.534
+DSoPo	$.460_{+1.08\%}$	$.655_{+1.55\%}$	$.756_{+0.93\%}$	$3.297_{+0.02\%}$	$9.925_{+0.033}$	$.495_{+7.30\%}$
+SoPo w/o VU	$.460_{+1.08\%}$	$.656_{+1.71\%}$	$.756_{+0.93\%}$	$3.295_{+0.02\%}$	$9.915_{+0.033}$	$.486_{+8.98\%}$
+USoPo	$.473_{+3.96\%}$	$.668_{+3.57\%}$	$.767_{+2.40\%}$	$3.226_{+2.36\%}$	$\mathbf{9.901}_{+0.047}$	$.556_{-4.12\%}$
+SoPo	$\mathbf{.479}_{+5.27\%}$	$\mathbf{.674}_{+4.50\%}$	$\mathbf{.770}_{+2.80\%}$	$\mathbf{3.208}_{+2.91\%}$	$9.906_{+0.042}$	$\mathbf{.480}_{+10.1\%}$
+SoPo ($\tau = 0.40$)	$.475_{+4.40\%}$	$.661_{+2.48\%}$	$.768_{+2.53\%}$	$3.272_{+0.97\%}$	$10.04_{-0.088}$	$.600_{-12.4\%}$
+SoPo ($\tau = 0.45$)	$\mathbf{.479}_{+5.27\%}$	$\mathbf{.674}_{+4.50\%}$	$\mathbf{.770}_{+2.80\%}$	$\mathbf{3.208}_{+2.91\%}$	$9.906_{+0.042}$	$.480_{+10.1\%}$
+SoPo ($\tau = 0.50$)	$.468_{+2.86\%}$	$.663_{+2.79\%}$	$.764_{+2.01\%}$	$3.256_{+1.45\%}$	$9.900_{+0.048}$	$.491_{+8.05\%}$
+SoPo ($\tau = 0.55$)	$.466_{+2.41\%}$	$.660_{+1.86\%}$	$.763_{+1.87\%}$	$3.263_{+1.24\%}$	$9.896_{+0.041}$	$.430_{+19.5\%}$
+SoPo ($\tau = 0.60$)	$.461_{+1.31\%}$	$.656_{+1.71\%}$	$.758_{+1.20\%}$	$3.288_{+0.48\%}$	$\mathbf{9.803}_{+0.145}$	$\mathbf{.399}_{+25.3\%}$
+SoPo ($K = 2$)	$\mathbf{.480}_{+5.50\%}$	$.671_{+4.03\%}$	$\mathbf{.771}_{+2.94\%}$	$3.212_{+2.78\%}$	$9.907_{+0.041}$	$.502_{+5.99\%}$
+SoPo ($K = 4$)	$.479_{+5.27\%}$	$\mathbf{.674}_{+4.50\%}$	$.770_{+2.80\%}$	$\mathbf{3.208}_{+2.91\%}$	$\mathbf{9.906}_{+0.042}$	$\mathbf{.480}_{+10.1\%}$

Table 4. **Ablation study on different alignment methods, cut-off thresholds τ, and numbers of sampled motions K.** The **best** results for each setting are in **bold**. The subscript indicates the magnitude of the performance increase or decrease.

of 5.34% (2.939 v.s. 3.096). Additionally, SoPo excels previous works, Motion Mamba and CrossDiff in primary metrics, including a Diversity advantage of 0.287 over Motion Mamba and an MM Dist improvement of 12.5% over CrossDiff. Furthermore, SoPo boosts FID for MLD* by 61.3%, highlighting its effectiveness. We also evaluate SoPo on the KIT-ML dataset, with results reported in Table 3. Integrated with MoMask [6], SoPo achieves the **best performance** in Top-k R-Precision (0.446, 0.673, 0.797), MM Dist (2.783), and FID (0.176). Furthermore, MLD integrated with SoPo outperforms the original one in all metrics. These results show the advantage of SoPo across various architectures.

5.2. Ablation Studies & Visualization

Impact of objective functions To evaluate the efficacy of SoPo, we fine-tuned the MDM model [28] using DSoPo (Eq. (12)), USoPo (Eq. (13)), SoPo without value-preferred motions (SoPo w/o VU), and SoPo (Eq. (14)). Results are reported in Table 4. DSoPo, trained directly on preference data comprising preferred motions from an offline dataset and unpreferred motions sampled online, partially mitigates limitations of online and offline DPO (Sec. 4.1), achieving a 7.30% FID improvement. By prioritizing preferred motions with reduced similarity to unpreferred ones, SoPo w/o VU enhances FID performance to 8.98%. Conversely, USoPo, with unpreferred motions filtered by threshold τ, improves Top 1 R-Precision, MM Dist, and Diversity by 3.96%, 2.36%, and 0.047, respectively, though FID decreases by 4.12%. A similar phenomenon also can be observed in the subsequent experiment, and we think that may be caused by the inappropriate parameter τ. Ultimately, SoPo integrates the strengths of these approaches, overcoming DSoPo's shortcomings and yielding optimal results: 5.27% in Top 1 R-Precision and 10.1% in FID.

Impact of different cut-off thresholds τ In Table 4, we vary cut-off thresholds τ from 0.40 to 0.60 in a step of 0.05 and present the corresponding results. The threshold τ determines the proportion of generated motions treated as preferred, with a larger τ value allowing more generated motions to be considered unpreferred. From Table 4, we can observe a universal pattern: as τ decreases, R-Precision and MM Dist are improved, while FID and Diversity are decreased. This can be explained by the fact that R-Precision and MM Dist are alignment metrics, and a cautious strategy with small τ that utilizes more accurate and valuable unpreferred motions for training can help the model improve the alignment performance. While FID and Diversity are metrics of generative quality, an aggressive strategy with a large τ that introduces more potential and uncertain unpreferred motions allows the model to avoid potential errors and achieve better generative quality.

Impact of different number of sample motions K Due to limitations in computational time and memory, we recommend keeping K below 8. In Table 4, we observe that generation quality improves significantly as K increases. This improvement stems from the expanded pool of generated motions, allowing the reward model to evaluate more samples. Consequently, the model better identifies and avoids unpreferred motions, leading to higher-quality generation.

6. Conclusion

In this study, we introduce a semi-online preference optimization method: a DPO-based fine-tune method for the text-to-motion model to directly align preference on "Semi-online data" consisting of high-quality preferred and diverse unpreferred motions. Our SoPo leverages the advantages both of online DPO and offline DPO, to overcome their own limitations. Furthermore, to ensure the validity of SoPo, we present a simple yet effective online generation method along with an offline reweighing strategy. Extensive experimental results show the effectiveness of our SoPo.

Limitation discussion. SoPo relies on a reward model to motion quality evaluation and identify usable unpreferred samples. However, research on reward models in the motion domain remains scarce, and current models, trained on specific datasets, exhibit limited generalization. Consequently, SoPo inherits these limitations, facing challenges in seamlessly fine-tuning diffusion models with reward models across diverse, open-domain scenarios.

References

[1] 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, 2, 6, 7, 8

[2] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. page 4302–4310. Advances in Neural Information Processing Systems. 3

[3] 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, 6, 7

[4] Ke Fan, Junshu Tang, Weijian Cao, Ran Yi, Moran Li, Jingyu Gong, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Freemotion: A unified framework for number-free text-to-motion synthesis. In *European Conference on Computer Vision*, pages 93–109. Springer, 2024. 2

[5] 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, 6, 7, 8

[6] 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, 8

[7] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024. 2

[8] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Advances in Neural Information Processing Systems*, pages 20067–20079. Curran Associates, Inc., 2023. 1, 2

[9] 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*, 36:20067–20079, 2023. 6, 7, 8

[10] Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, and Xinchao Wang. Priority-Centric Human Motion Generation in Discrete Latent Space . In *IEEE/CVF International Conference on Computer Vision*, pages 14760–14770, Los Alamitos, CA, USA, 2023. IEEE. 1, 2, 6, 7

[11] Shuai Li, Sisi Zhuang, Wenfeng Song, Xinyu Zhang, Hejia Chen, and Aimin Hao. Sequential texts driven cohesive motions synthesis with natural transitions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9498–9508, 2023. 2

[12] 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

[13] Zhanhao Liang, Yuhui Yuan, Shuyang Gu, Bohan Chen, Tiankai Hang, Ji Li, and Liang Zheng. Step-aware preference optimization: Aligning preference with denoising performance at each step. *arXiv preprint arXiv:2406.04314*, 2024. 2

[14] Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, and Chang-wen Chen. Being comes from not-being: Open-vocabulary text-to-motion generation with wordless training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23222–23231, 2023. 2

[15] Zichen Miao, Zhengyuan Yang, Kevin Lin, Ze Wang, Zicheng Liu, Lijuan Wang, and Qiang Qiu. Tuning timestep-distilled diffusion model using pairwise sample optimization, 2024. 2

[16] Sanghyeon Na, Yonggyu Kim, and Hyunjoon Lee. Boost your own human image generation model via direct preference optimization with ai feedback. *ArXiv*, abs/2405.20216, 2024. 2

[17] Massimiliano Pappa, Luca Collorone, Giovanni Ficarra, Indro Spinelli, and Fabio Galasso. Modipo: text-to-motion alignment via ai-feedback-driven direct preference optimization, 2024. 1, 2, 4, 6, 7

[18] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, 2022. 6, 7, 8

[19] Mathis Petrovich, Michael J. Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9454–9463, 2023. 6

[20] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. MMM: Generative Masked Motion Model . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555. IEEE, 2024. 1, 2

[21] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2): 193–202, 1975. 3

[22] Matthias Plappert, Christian Mandery, and Tamim Asfour. The kit motion-language dataset. *Big Data*, 4(4):236–252, 2016. PMID: 27992262. 6

[23] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2018. 2

[24] Qiaosong Qi, Le Zhuo, Aixi Zhang, Yue Liao, Fei Fang, Si Liu, and Shuicheng Yan. Diffdance: Cascaded human motion diffusion model for dance generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 1374–1382, New York, NY, USA, 2023. Association for Computing Machinery. 1

[25] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Advances in Neural Information Process-*

ing Systems, Red Hook, NY, USA, 2024. Curran Associates Inc. 2, 3, 6

[26] Zeping Ren, Shaoli Huang, and Xiu Li. Realistic human motion generation with cross-diffusion models. *European Conference on Computer Vision*, 2024. 1, 2, 6, 7

[27] Yoni Shafir, Guy Tevet, Roy Kapon, and Amit Haim Bermano. Human motion diffusion as a generative prior. In *The Twelfth International Conference on Learning Representations*, 2024. 6, 7

[28] 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, 6, 7, 8

[29] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2, 6

[30] Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, and Xiaohui Liang. Fg-t2m: Fine-grained text-driven human motion generation via diffusion model. In *IEEE/CVF International Conference on Computer Vision*, pages 21978–21987, 2023. 1, 6, 7

[31] Zan Wang, Yixin Chen, Baoxiong Jia, Puhao Li, Jinlu Zhang, Jingze Zhang, Tengyu Liu, Yixin Zhu, Wei Liang, and Siyuan Huang. Move as you say, interact as you can: Language-guided human motion generation with scene affordance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 433–444, 2024. 1, 6, 7

[32] Kai Yang, Jian Tao, Jiafei Lyu, Chunjiang Ge, Jiaxin Chen, Weihan Shen, Xiaolong Zhu, and Xiu Li. Using Human Feedback to Fine-tune Diffusion Models without Any Reward Model . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8941–8951, Los Alamitos, CA, USA, 2024. IEEE Computer Society. 2

[33] Junliang Ye, Fangfu Liu, Qixiu Li, Zhengyi Wang, Yikai Wang, Xinzhou Wang, Yueqi Duan, and Jun Zhu. Dreamreward: Text-to-3d generation with human preference. *arXiv preprint arXiv:2403.14613*, 2024. 2

[34] Daoan Zhang, Guangchen Lan, Dong-Jun Han, Wenlin Yao, Xiaoman Pan, Hongming Zhang, Mingxiao Li, Pengcheng Chen, Yu Dong, Christopher Brinton, and Jiebo Luo. Seppo: Semi-policy preference optimization for diffusion alignment, 2024. 2

[35] 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. 8

[36] 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*, pages 14730–14740, 2023. 2

[37] 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

[38] 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. 1, 2, 6, 7, 8

[39] 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

[40] 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. 1, 6, 7, 8

[41] Zixiang Zhou and Baoyuan Wang. Ude: A unified driving engine for human motion generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5632–5641, 2023. 2

[42] Banghua Zhu, Michael Jordan, and Jiantao Jiao. Iterative data smoothing: Mitigating reward overfitting and overoptimization in RLHF. In *Proceedings of the 41st International Conference on Machine Learning*, pages 62405–62428. PMLR, 2024. 3

[43] Wentao Zhu, Xiaoxuan Ma, Dongwoo Ro, Hai Ci, Jinlu Zhang, Jiaxin Shi, Feng Gao, Qi Tian, and Yizhou Wang. Human motion generation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2