Direct Preference Optimization for Text-to-Motion Generation with Noisy AI Feedback

Yi-Hsuan Lin
R13944021@ntu.edu.tw

Chi-Jun You
R14921a07@ntu.edu.tw

Yan-Ting Chen
B11303039@ntu.edu.tw

Abstract

This work investigates the viability and limitations of Reinforcement Learning from AI Feedback (RLAIF) in Text-to-Motion (T2M) alignment. We uncover a counter-intuitive "negative scaling" phenomenon: while DPO improves performance on a small-scale curated dataset ($\sim$2.2k), scaling to a large synthetic dataset (14.6k) degrades alignment quality, suggesting that systematic inherent noise in AI labelers accumulates to disrupt learning. To quantify this vulnerability, we conduct a systematic stress test using label flipping noise. Our results demonstrate that DPO exhibits notable robustness to random noise (maintaining superiority over the baseline up to a 60% noise ratio); however, training collapses under adversarial settings, underscoring that the reliability of the AI Oracle is the critical bottleneck for scaling preference learning.

1 Introduction

Recent advancements in generative models have significantly propelled the field of Text-to-Motion (T2M) generation. Despite the success of diffusion-based approaches, aligning generated motions with fine-grained human preferences remains a challenge. Standard metrics (e.g., FID, R-Precision) do not always correlate with human aesthetic judgment.

While Reinforcement Learning from Human Feedback (RLHF) has proven effective in Large Language Models (LLMs), collecting large-scale human preference data for 3D motion is prohibitively expensive and time-consuming. This motivates the exploration of **RL from AI Feedback (RLAIF)**, utilizing advanced AI models as evaluators. However, AI annotators are not infallible; they may hallucinate or misinterpret prompts, introducing "inherent noise" into the training signal that can mislead the optimization process.

In this work, we investigate the feasibility and limitations of AI-assisted Direct Preference Optimization (DPO) for T2M. Unlike previous works that assume better performance with more AI data, we uncover a critical trade-off between data scale and label reliability. Our contributions are three-fold:

1. We establish a DPO training pipeline for motion generation using 14.6k AI-generated preference pairs derived from the VimoRAG reward model.
2. We conduct a critical evaluation of RLAIF scaling, revealing a **negative scaling** phenomenon: while DPO improves alignment on a small-scale dataset ($\sim$2.2k), performance degrades significantly when scaling to the full synthetic dataset (14.6k), suggesting that systematic AI labeling noise accumulates to disrupt learning.
3. We perform a systematic stress test by injecting *Label Flipping Noise*, empirically identifying that DPO in the motion domain is robust to random noise (maintaining superiority over the baseline up to $\epsilon \leq 0.6$) but suffers catastrophic collapse under adversarial settings ($\epsilon = 1.0$), quantifying the breakdown threshold for future RLAIF systems.

Preprint.

2 Related Works

2.1 Text-to-Motion Generation.

Early approaches to human motion synthesis primarily employed VAEs and GANs. However, the field has recently shifted towards diffusion probabilistic models, which have demonstrated superior capability in synthesizing diverse and high-fidelity human motions. In this work, we build upon MotionGPT [1], a prominent codebook-based generative model that treats motion generation as a language modeling task. We adopt MotionGPT as our backbone due to its robust performance and established status as a baseline in the T2M domain.

2.2 Preference Optimization in Motion Synthesis.

Direct Preference Optimization (DPO), proposed by [2], has emerged as a stable and efficient alternative to PPO-based Reinforcement Learning from Human Feedback (RLHF). Unlike PPO, DPO optimizes the policy directly from preference data without the need to train a separate, often unstable, reward model. In the motion domain, the standard benchmark is HumanML3D [3], which contains 14,616 text-motion pairs derived from real human movements. Recent works have begun to explore preference alignment in T2M. [4] pioneered the use of human-labeled data to fine-tune T2M models, creating an additional 3,528 preference pairs. Their comparative analysis indicates that DPO outperforms RLHF+PPO, hypothesizing that the paucity of preference data makes it difficult to train a reliable reward model required for PPO. Furthermore, VimoRAG [5] integrates Retrieval-Augmented Generation (RAG) with motion synthesis. It retrieves video clips to serve as prompt prefixes and employs a specialized AI-based dual-aligned reward model to facilitate a two-stage McDPO fine-tuning process.

$$r(x, v, \hat{y}_i) = -(w_l \frac{l(\hat{y}_i, y)}{\sum_{j \in k} l(\hat{y}_j, y)} + w_d \frac{d(\hat{y}_i, x)}{\sum_{j \in k} d(\hat{y}_j, x)})$$

where y represents the reference motion derived from the retrieved video v, $\hat{y}_i$ denotes the i-th generated motion candidate, and x is the input text prompt. The hyperparameters w_l and w_d control the weights for the motion-to-motion reconstruction loss $l(\cdot)$ and the text-to-motion alignment distance $d(\cdot)$, respectively, normalized over the candidate set k.

2.3 Learning with Noisy Labels.

The impact of noisy labels is a well-studied problem in classification but remains relatively under-explored in the context of preference alignment, particularly outside the NLP domain. In the realm of Large Language Models (LLMs), [6] systematically analyzed the effect of preference noise on alignment performance. They observed that alignment remains beneficial when noise rates are below 30%; however, performance degrades significantly as noise exceeds 40%, and the training signal becomes detrimental as noise approaches 50%. Despite these insights in NLP, the resilience of Text-to-Motion models to such feedback noise remains unexamined. To the best of our knowledge, our work is the first to specifically address the scenario of "label flipping" in T2M preference optimization, investigating the robustness of DPO against AI-induced annotation errors.

3 Problem Formulation

We formulate the text-to-motion alignment task as a reinforcement learning problem, specifically treating conditional generation as a contextual bandit problem.

RL Framework Definition. Let $\mathcal{X}$ denote the space of text prompts and $\mathcal{Y}$ the continuous space of 3D motion sequences. We define the RL components as follows:

- **State (s):** The context provided by the input text description $x \in \mathcal{X}$.
- **Action (a):** The generated motion sequence $y \in \mathcal{Y}$.
- **Policy (π_θ):** The generative model parameterized by θ, which maps the state x to a probability distribution over actions y, denoted as $\pi_\theta(y|x)$.

- **Reward (r):** A latent reward function $r^*(x, y)$ that reflects the alignment between the generated motion and human preference.

Optimization Objective. Our goal is to learn an optimal policy π^* that maximizes the expected reward while remaining close to the reference policy π_{ref} (the supervised fine-tuned model) to maintain generation diversity and stability. This corresponds to the standard RLHF objective:

$$\max_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[r^*(x, y) - \beta\mathbb{D}_{KL}\left(\pi_\theta(y|x)||\pi_{ref}(y|x)\right)\right] \tag{1}$$

where β is a coefficient controlling the strength of the KL-divergence penalty.

Derivation for DPO. Traditionally, this objective is solved by training a reward model to approximate r^* and then optimizing π_θ via PPO. However, Direct Preference Optimization (DPO) derives the analytical solution for the optimal policy π^* in terms of the reward function. By rearranging the terms, the reward can be expressed as a function of the policy ratio. This allows us to optimize the policy directly using preference data pairs (y_w, y_l) without an explicit reward model, effectively bypassing the unstable actor-critic training loop.

4 Method

In this section, we outline our framework for aligning text-to-motion models using AI feedback. We first introduce the mathematical formulation of Direct Preference Optimization (DPO). Next, we detail our pipeline for generating preference data using an AI oracle. Finally, we describe our noise injection mechanism and experimental design to stress-test the robustness of DPO against label flipping.

4.1 Preliminaries: Direct Preference Optimization

We formulate the text-to-motion alignment problem as optimizing a policy π_θ to align with a preference distribution. Given a text prompt x and a pair of generated motions (y_w, y_l), where y_w is preferred over y_l, DPO optimizes the policy by minimizing the negative log-likelihood of the preference data. Unlike PPO, which requires an explicit reward model, DPO implicitly optimizes the reward by solving:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right] \tag{2}$$

where π_{ref} is the frozen reference model (in our case, the supervised fine-tuned MotionGPT), σ is the sigmoid function, and β is a hyperparameter controlling the divergence from the reference policy.

4.2 RLAIF Data Pipeline

To scale up preference learning without expensive human annotation, we establish a Reinforcement Learning from AI Feedback (RLAIF) pipeline. We utilize **MotionGPT3** [7] (instead of the original MotionGPT [1]) as our generator to synthesize 14,616 motion pairs from the HumanML3D training prompts, aiming to mitigate potential environmental instability issues.

To assign preference labels, we employ the reward model from VimoRAG [5] as our AI Oracle. Specifically, the oracle evaluates a generated motion $\hat{y}_i$ given the text prompt x and the retrieved video reference v using a dual-aligned reward function defined as:

$$r(x, v, \hat{y}_i) = -\left(w_l\frac{l(\hat{y}_i, y)}{\sum_{j\in k} l(\hat{y}_j, y)} + w_d\frac{d(\hat{y}_i, x)}{\sum_{j\in k} d(\hat{y}_j, x)}\right) \tag{3}$$

where $l(\cdot)$ represents the reconstruction loss measuring motion quality, and $d(\cdot)$ represents the feature distance measuring text-motion alignment. The terms are normalized across the candidate set k, and w_l, w_d are weighting factors. Due to the absence of ground-truth reference motions y in the

InstructMotion [4] dataset (which prevents the calculation of the reconstruction loss $l(\cdot)$), we set $w_l = 0$ and $w_d = 1$, thereby evaluating preferences based exclusively on text-motion alignment.

Based on this calibrated reward score, we rank the candidate pairs such that $r(y_w) > r(y_l)$ to construct our synthetic preference dataset $\mathcal{D}_{AI}$.

4.3 Noise Injection and Experimental Design

Quantifying Inherent AI Noise. While the RLAIF pipeline enables scalable data generation, the reliability of the AI Oracle remains a critical concern. To quantify the discrepancy between AI judgments and human intuition, we conducted a preliminary pilot study using the human-annotated dataset from [4]. This dataset originally contains 3,528 pairs; however, we filtered out 1,312 pairs labeled as "skipped" (where both motions were deemed low quality), resulting in a high-quality subset of 2,216 pairs.

We evaluated these ground-truth pairs using our VimoRAG reward model. This pilot test revealed a **34% disagreement rate** between the AI's rankings and human labels. We define this 34% gap as *inherent preference noise*.

Stress Test with Label Flipping. To systematically investigate how such preference noise impacts DPO training stability, we design a controlled stress test. We treat the 2,216 human-labeled pairs as our noise-free baseline $\mathcal{D}_{clean}$. We then construct noisy datasets $\tilde{\mathcal{D}}_\epsilon$ by injecting *Label Flipping Noise*. For each pair (y_w, y_l), we flip the preference direction with probability ϵ:

$$(y'_w, y'_l) = \begin{cases} (y_l, y_w) & \text{if } u < \epsilon \\ (y_w, y_l) & \text{otherwise} \end{cases} \tag{4}$$

where $u \sim \text{Uniform}(0, 1)$. We experiment with three specific noise levels using five different noise ratios:

- $\epsilon \in \{0.05, 0.2\}$: Simulating scenarios with high-quality but imperfect annotation (typical human error or high-performing AI).
- $\epsilon = 0.4$: Simulating the approximate noise level observed in our pilot study (close to the 34% gap), representing a realistic noisy AI annotator.
- $\epsilon \in \{0.6, 1.0\}$: Simulating adversarial settings where the majority of labels are incorrect (worse than random guessing) to test the breakdown threshold.

This design allows us to isolate the noise factor and empirically determine the robustness of the DPO algorithm in the motion domain.

5 Experimental Results

5.1 Experimental Setup

Dataset. We conduct our experiments on the **HumanML3D** dataset [3], a standard benchmark for text-to-motion generation containing 14,616 motion clips with corresponding textual descriptions. **Implementation Details.** We use **MotionGPT** [1] as our reference model (SFT Base). For DPO training, we set the learning rate to $1e-5$, $\beta = 0.1$, and use a batch size of 64. The AI feedback oracle is the reward model from VimoRAG [5]. **Evaluation Metrics.** We evaluate performance using three standard metrics: **Evaluation Metrics.** We report the following metrics consistent with standard benchmarks:

- **R-Precision** (Top-1, 2, 3) (↑): Measures the retrieval accuracy of the ground-truth text description given the generated motion. We report Top-1, Top-2, and Top-3 accuracy.
- **Multi-Modality (MModality)** (↑): Measures the diversity/richness of motions generated from the same text prompt (higher is better).
- **FID** (↓): Frechet Inception Distance, measuring the distance between the distribution of generated motions and real motions.

- **Diversity** ($\rightarrow$): Measuring the variance of the generated motions across the dataset (values closer to the real data distribution are preferred).

5.2 Results

5.2.1 Effectiveness of Scaling AI Feedback

We evaluate the impact of dataset scale and label source on alignment performance. We conducted two distinct training runs using AI-generated preferences:

1. **AI-DPO (2.2k):** We utilized the 2,216 motion pairs from the InstructMotion dataset but replaced the ground-truth human labels with our AI Oracle's judgments. This allows for a direct comparison with the Human-DPO baseline on identical motion candidates.
2. **AI-DPO (14.6k):** We utilized our self-generated dataset of 14,616 pairs derived from HumanML3D prompts, labeled by the same AI Oracle.

As presented in Table 1, the results highlight two critical insights:

(1) The Cost of AI Noise (Human vs. AI). Comparing the models trained on the same 2.2k motion pairs, the **Human-DPO** baseline (0.425) outperforms **AI-DPO** (0.413). Since the motion candidates are identical, this performance drop is solely attributable to the inaccuracy of the AI labeler compared to human experts. This confirms the existence of *inherent preference noise* in the VimoRAG reward model.

(2) Negative Scaling (Small vs. Large). Contrary to the standard scaling law where "more data yields better performance," scaling up to 14.6k synthetic pairs causes performance to degrade further to 0.390. This indicates that the AI Oracle's label noise is not random but systematic; simply adding more AI-labeled data accumulates misleading gradients rather than correcting them. This failure to scale directly motivates our subsequent stress test to quantify the noise tolerance of DPO.

Table 1: Impact of Label Source and Data Scale. Note: **M-Mod**: MModality, **Div**: Diversity.

		Alignment			**Quality**		
Method	Scale	Top-1↑	Top-2↑	Top-3↑	M-Mod↑	FID↓	Div→
MotionGPT (Base)	-	0.405	0.567	0.658	**3.495**	**0.178**	9.393
Human-DPO (Gold)	2.2k	**0.425**	**0.604**	**0.704**	2.264	0.266	9.592
AI-DPO (Re-labeled)	2.2k	0.413	0.591	0.689	2.503	0.233	9.661
AI-DPO (Synthetic)	14.6k	0.390	0.548	0.641	2.801	0.237	9.556

5.2.2 Impact of Label Flipping Noise

To analyze the robustness of the DPO algorithm against preference noise, we conducted controlled experiments on the human-labeled subset by injecting label flipping noise at rates $\epsilon \in \{0, 0.05, 0.2, 0.6\}$.

Table 2 and figure 1 summarize the results. We observe that:

1. **Robustness at low noise ($\epsilon \leq 0.05$):** The performance remains comparable to the noise-free baseline ($\epsilon = 0$), indicating DPO is resilient to minor annotation errors inherent in human or high-quality AI feedback.
2. **Degradation at moderate noise ($\epsilon = 0.2, 0.4$):** While there is a drop in FID and R-Precision, the model still outperforms the SFT Base, suggesting DPO can extract useful signals even from moderately noisy data.
3. **Failure at adversarial noise ($\epsilon = 1.0$):** When the majority of labels are flipped (worse than random guessing), the optimization collapses, leading to performance significantly worse than the base model. This confirms the necessity of maintaining a noise rate higher than a certain rate for effective alignment.

Table 2: Robustness analysis under label flipping noise. **M-Mod**: MModality, **Div**: Diversity.

		Alignment			**Quality**		
Method	Noise	Top-1↑	Top-2↑	Top-3↑	M-Mod↑	FID↓	Div→
MotionGPT	-	0.405	0.567	0.658	3.495	**0.178**	9.393
DPO (Clean)	**0**	**0.425**	**0.604**	**0.704**	2.264	0.266	9.592
DPO (Minor)	0.05	0.421	0.600	0.700	2.328	0.237	9.564
DPO (High)	0.2	0.419	0.597	0.695	2.459	0.227	9.505
DPO (High)	0.4	0.418	0.589	0.682	2.229	0.189	9.601
DPO (Adv.)	0.6	0.409	0.577	0.670	2.614	0.295	9.463
DPO (Adv.)	1.0	0.372	0.529	0.617	**3.531**	0.344	9.350

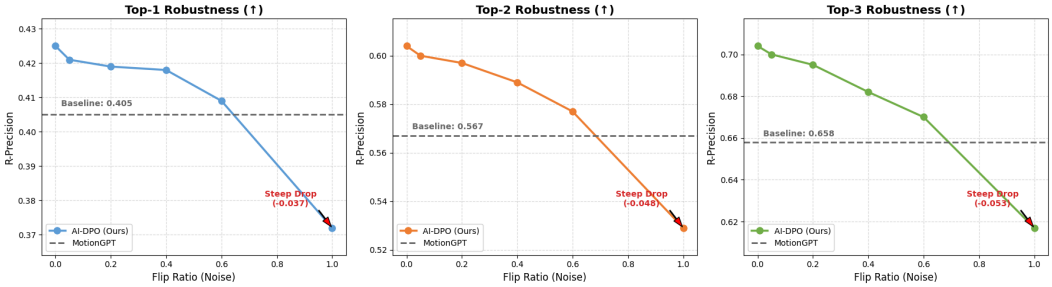

Figure 1: Impact of Label Flipping Noise on Text-Motion Alignment. We plot R-Precision (Top-1, Top-2, Top-3) against increasing noise rates (ϵ). The concave degradation curve indicates that DPO is relatively robust to low-level noise (maintaining performance above the MotionGPT baseline up to $\epsilon \approx 0.6$) but suffers accelerating performance decay as noise increases, culminating in a steep drop at adversarial levels ($\epsilon = 1.0$)

6 Conclusion and Future Work

In this work, we investigated the boundaries of AI-assisted Text-to-Motion alignment. We demonstrated that while RLAIF is feasible, its efficacy is constrained by the reliability of the reward signal. Our systematic stress test with label flipping reveals a concave degradation pattern: DPO exhibits remarkable resilience to random annotation errors, maintaining performance superiority over the base model even when noise rates reach as high as 60% ($\epsilon = 0.6$). However, the method is not invincible; performance decays acceleratingly as noise increases, culminating in catastrophic collapse under adversarial settings ($\epsilon = 1.0$).

Limitations. The primary bottleneck identified is the *quality*, rather than the quantity, of the AI feedback. As quantified by our pilot study, the VimoRAG oracle suffers from a 34% inherent disagreement rate with human annotators. While our stress test proves DPO can handle this *volume* of random noise, the *systematic* nature of AI hallucinations likely imposes a harder ceiling on alignment performance than random label flipping. Additionally, our study is limited to a single generator (MotionGPT) and dataset (HumanML3D), leaving the cross-architecture generalizability of these noise thresholds an open question.

Future Work. To overcome the inherent noise barrier, future research should move beyond standard DPO. We recommend investigating noise-robust objectives such as **rDPO** (Robust DPO) or **ROPO** (Robust Preference Optimization), which explicitly model and downweight noisy samples. Furthermore, a hybrid "Human-in-the-loop" approach—using a small set of gold-standard human labels to calibrate the AI Oracle's confidence—could effectively filter out the systematic errors that currently hinder scaling.

References

[1] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language, 2023.

[2] 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, 2024.

[3] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

[4] Jenny Sheng, Matthieu Lin, Andrew Zhao, Kevin Pruvost, Yu-Hui Wen, Yangguang Li, Gao Huang, and Yong-Jin Liu. Exploring text-to-motion generation with human preference, 2024.

[5] Haidong Xu, Guangwei Xu, Zhedong Zheng, Xiatian Zhu, Wei Ji, Xiangtai Li, Ruijie Guo, Meishan Zhang, Min zhang, and Hao Fei. Vimorag: Video-based retrieval-augmented 3d motion generation for motion language models, 2025.

[6] Yang Gao, Dana Alon, and Donald Metzler. Impact of preference noise on the alignment performance of generative language models, 2024.

[7] Bingfan Zhu, Biao Jiang, Sunyi Wang, Shixiang Tang, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. Motiongpt3: Human motion as a second modality, 2025.