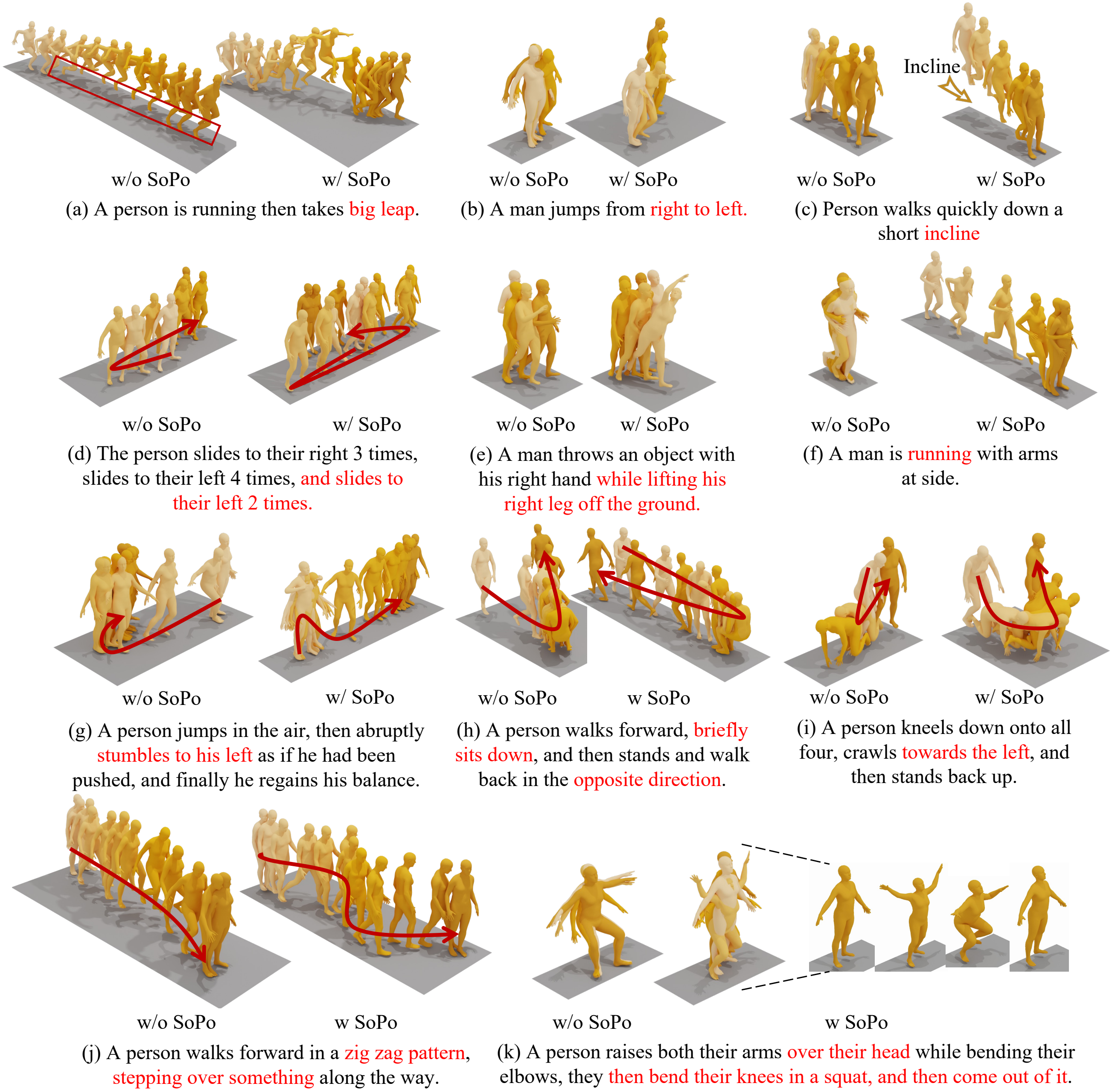
SoPo: Text-to-Motion Generation Using Semi-Online Preference Optimization

Supplementary Material



(a) A person is running then takes big leap.

(b) A man jumps from right to left.

(c) Person walks quickly down a short incline

(d) The person slides to their right 3 times, slides to their left 4 times, and slides to their left 2 times.

(e) A man throws an object with his right hand while lifting his right leg off the ground.

(f) A man is running with arms at side.

(g) A person jumps in the air, then abruptly stumbles to his left as if he had been pushed, and finally he regains his balance.

(h) A person walks forward, briefly sits down, and then stands and walk back in the opposite direction.

(i) A person kneels down onto all four, crawls towards the left, and then stands back up.

(j) A person walks forward in a zig zag pattern, stepping over something along the way.

(k) A person raises both their arms over their head while bending their elbows, they then bend their knees in a squat, and then come out of it.

Figure S1. Visual results on HumanML3D dataset. We integrate our SoPo into 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.

This supplementary document contains the technical proofs of results and some additional experimental results. It is structured as follows. Sec. A provides the implementation and theoretical analysis of our SoPo. Sec. B gives the proofs of the main results, including Theorem 1, Theorem 2, the objective function of DSoPo, the objective function of USoPo, and theorem of SoPo for text-to-motion generation. Then in Sec. C presents the additional experiment information, including additional experimental details (Sec. C.1) and results (Sec. C.2).

A. Details of SoPo for Text-to-Motion Generation

In this section, we first examine the objective function of SoPo and argue that it presents significant challenges for optimization. Fortunately, we then discover and derive an equivalent form that is easier to optimize (Sec. A.1). Finally, we design an algorithm to optimize it and finish discussing their correspondence (Sec. A.2).

A.1. Equivalent form of SoPo

In Eq. (15) and (16), the objective function of SoPo is defined as:

$$\mathcal{L}^{\text{diff}}_{\text{SoPo}} = \mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{vu}} + \mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{hu}}, \tag{S1}$$

$$\begin{aligned}\mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{vu}} &= -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D},x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}^{vu*}_\theta(\cdot|c)} Z_{vu}(c)\Big[\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big)\Big]\\ \mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{hu}} &= -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}} Z_{hu}(c)\Big[\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\text{ref},x^w_t)\Big)\Big]\end{aligned} \tag{S2}$$

However, these objectives can not be directly optimized, since the distribution $\bar{\pi}^{vu*}_\theta$ and $\bar{\pi}^{hu*}_\theta$ are not defined explicitly. To this end, we begin by inducing its equivalent form:

$$\mathcal{L}^{\text{diff}}_{\text{SoPo}}(\theta) = -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D},x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}_\theta(\cdot|c)}\begin{cases}\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big), & \text{If } r(x^l,c)<\tau,\\ \log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\text{ref},x^w_t)\Big), & \text{Otherwise.}\end{cases} \tag{S3}$$

where $x^l = \text{argmin}_{\{x^k_{\bar{\pi}_\theta}\}^K_{k=1}\sim\pi_\theta} r(x^k_{\pi_\theta}, c)$.

Proof. Recall our definition of $\mathcal{L}^{\text{diff}}_{\text{SoPo}}(\theta)$ in Eq. (15) and (16). Through algebraic maneuvers, we have:

$$\begin{aligned}\mathcal{L}^{\text{diff}}_{\text{SoPo}} =\; & \mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{vu}} + \mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{hu}}\\ =& -\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,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big)\Big]\\ & -\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,\text{ref},x^w_t)\Big)\Big]\\ =& -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}\mathbb{E}_{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,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big)\Big]\\ & -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}^{hu*}_\theta(\cdot|c)} Z_{hu}(c)\Big[\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\text{ref},x^w_t)\Big)\Big]\\ =& -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}}\underbrace{p^{vu*}_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)Z_{vu}(c)}_{p^{vu}_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)}\Big[\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big)\Big]\\ & -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}}\underbrace{p^{hu*}_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)Z_{hu}(c)}_{p^{hu}_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)}\Big[\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\text{ref},x^w_t)\Big)\Big]\\ \overset{①}{=}& -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}_\theta(\cdot|c)}p_\tau(r(x^l_{\bar{\pi}_\theta},c)<\tau)\Big[\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big)\Big]\\ & -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}_\theta(\cdot|c)}p_\tau(r(x^l_{\bar{\pi}_\theta},c)\geq\tau)\Big[\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\text{ref},x^w_t)\Big)\Big]\\ =& -\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,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{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,\text{ref},x^w_t)\Big), & \text{Otherwise.}\end{cases}\end{aligned}$$

where ① holds since $p_{\bar{\pi}^{vu*}_\theta}(\cdot) = \frac{p^{vu}_{\bar{\pi}_\theta}(\cdot)}{Z_{vu}(c)}$ and $p^{vu}_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c) = p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)\cdot p_\tau(r(x^l_{\bar{\pi}_\theta},c)\geq\tau)$.The proof is completed. □

A.2. The process of SoPo for text-to-motion generation

Based on the equivalent form of SoPo in Eq. (S15), we can design an algorithm to directly optimize it, as shown in **Algorithm 1**.

Algorithm 1 SoPo for text-to-motion generation

```
Input:Preference datasets with only preferred motions $(x^w, c) \in \mathcal{D}$; Number of diffusion steps $T$; Number of iterations $I$; Number of generated motions $K$; Reference model $\pi_{\text{ref}}$; Policy model $\pi_\theta$; Cut-off threshold value $\tau$
Output: The aligned model $\pi_\theta$;
 1: for $i = 1, 2, .., I$ do
 2:     for each $(x^w, c) \in \mathcal{D}$ do
 3:         Sample diffusion step $t \sim \mathcal{U}(0, T)$
 4:         Sample $K$ motions $x^{1:K}_{\bar{\pi}_\theta}$ from $\bar{\pi}_\theta(\cdot|c)$
 5:         Compute the weight of the preferred motion $x^w$: $S(x^w) = \min_{x^k_{\bar{\pi}_\theta} \sim \pi_\theta} \cos(x^w, x^k_{\bar{\pi}_\theta})$
 6:         Select the unpreferred motion with the lowest preference score: $x^l = \text{argmin}_{\{x^k_{\bar{\pi}_\theta}\}^K_{k=1} \sim \pi_\theta} r(x^k_{\pi_\theta}, c)$
 7:         if $r(x^l, c) < \tau$ then
 8:             $\mathcal{L}(\theta) = \log \sigma\Big( - T\omega_t \big(\beta_w(x_w)(\mathcal{L}(\theta, \text{ref}, x^w_t) - \beta\mathcal{L}(\theta, \text{ref}, x^l_t))\Big)$
 9:         else
10:             $\mathcal{L}(\theta) = \log \sigma\Big( - T\omega_t \beta_w(x_w)(\mathcal{L}(\theta, \text{ref}, x^w_t)\Big)$
11:         end if
12:         $\mathcal{L}^{\text{diff}}_{\text{SoPo}}(\theta) = \mathcal{L}^{\text{diff}}_{\text{SoPo}}(\theta) + \mathcal{L}(\theta)$
13:     end for
14:     Update policy model $\pi_\theta$ by $\nabla_\theta \mathcal{L}^{\text{diff}}_{\text{SoPo}}(\theta)$
15: end for
16: return The aligned policy model $\pi_\theta$
```

The SoPo optimizes a policy model π_θ for text-to-motion generation through an iterative process guided by a reward model. In each iteration, given a preferred motion x^w and a conditional code c, a random diffusion step t is selected, and K candidate motions are generated by π_θ. The motion with the lowest preference score is then treated as the unpreferred motion. To determine the weight of the preferred motion x^w, the similarities between all generated motions are computed, and the lowest cosine similarity value is used to calculate its weight. Finally, the loss is calculated in two ways, determined based on the preference scores of the unpreferred motion. If the preference score of the selected unpreferred motion falls below a threshold τ, it is identified as a valuable unpreferred motion and used for training. Otherwise, it indicates that the motions generated by the policy model π_θ are satisfactory. In such cases, the policy model is trained exclusively on high-quality preferred motions, rather than on both preferred motions and relatively high-preference unpreferred motions.

To further understand the objective function, we analyze the correspondence between the objective function in Eq. (S15) and Algorithm 1:

$$\mathcal{L}^{\text{diff}}_{\text{SoPo}}(\theta) = -\mathbb{E}_{\underbrace{(x^w, c) \sim \mathcal{D}}_{\text{Line 2}}, \underbrace{t \sim \mathcal{U}(0,T)}_{\text{Line 3}}, \underbrace{x^{1:K}_{\bar{\pi}_\theta} \sim \bar{\pi}_\theta(\cdot|c)}_{\text{Line 4}}} \begin{cases} \underbrace{\log \sigma\Big(- T\omega_t \big(\beta_w(x_w)(\mathcal{L}(\theta, \text{ref}, x^w_t) - \beta\mathcal{L}(\theta, \text{ref}, x^l_t))\Big)}_{\text{Line 8}}, & \underbrace{\text{If } r(x^l, c) < \tau,}_{\text{Line 7}} \\ \underbrace{\log \sigma\Big(- T\omega_t \beta_w(x_w)\mathcal{L}(\theta, \text{ref}, x^w_t)\Big)}_{\text{Line 10}}, & \underbrace{\text{Otherwise}}_{\text{Line 9}}. \end{cases} \tag{S4}$$

B. Theories

B.1. Proof of Theorem 1

Proof. The offline DPO based on Plackett-Luce model [7] can be denoted as:

$$\mathcal{L}_{\text{off}}(\theta) = -\mathbb{E}_{(x^{1:K}, c) \sim \mathcal{D}} \Big[\log \prod_{k=1}^{K} \frac{\exp(\beta h_\theta(x^k, c))}{\sum_{j=k}^{K} \exp(\beta h_\theta(x^j, c))} \Big], \tag{S5}$$

where $h_\theta(x,c) = \log \frac{\pi_\theta(x|c)}{\pi_{\text{ref}}(x|c)}$. Then we have:

$$\begin{aligned}
\mathcal{L}_{\text{off}}(\theta) &= -\mathbb{E}_{(x^{1:K},c)\sim\mathcal{D}}\Big[\log\prod_{k=1}^{K}\frac{\exp(\beta h_\theta(x^k,c))}{\sum_{j=k}^{K}\exp(\beta h_\theta(x^j,c))}\Big] \\
&= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log\prod_{k=1}^{K}\frac{\exp(\beta h_\theta(x^k,c))}{\sum_{j=k}^{K}\exp(\beta h_\theta(x^j,c))}\Big] \\
&= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log\prod_{k=1}^{K}\frac{\exp(\beta\log\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)})}{\sum_{j=k}^{K}\exp(\beta\log\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)})}\Big] \\
&= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log\prod_{k=1}^{K}\frac{\exp\log(\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)})^\beta)]}{\sum_{j=k}^{K}\exp\log(\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)})^\beta}\Big] \\
&= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log\prod_{k=1}^{K}\underbrace{\frac{(\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)})^\beta}{\sum_{j=k}^{K}(\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)})^\beta}}_{p_\theta(x^k|c)}\Big] \\
&= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log\underbrace{\prod_{k=1}^{K}p_\theta(x^k|c)}_{p_\theta(x^{1:K}|c)}\Big] \\
&= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log p_\theta(x^{1:K}|c) - \log p_{\text{gt}}(x^{1:K}|c) + \log p_{\text{gt}}(x^{1:K}|c)\Big] \\
&= \mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; p_{\text{gt}}(x^{1:K}|c)\Big[\log\frac{p_{\text{gt}}(x^{1:K}|c)}{p_\theta(x^{1:K}|c)} - \log p_{\text{gt}}(x^{1:K}|c)\Big] \\
&= \mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\; D_{KL}(p_{\text{gt}}|p_\theta) - p_{\text{gt}}(x^{1:K}|c)\log p_{\text{gt}}(x^{1:K}|c)
\end{aligned} \tag{S6}$$

Therefore, we have:

$$\nabla_\theta\mathcal{L}_{\text{off}}(\theta) = \mathbb{E}_{c\sim\mathcal{D},x^{1:K}}\nabla_\theta D_{KL}(p_{\text{gt}}||p_\theta). \tag{S7}$$

The proof is completed. □

B.2. Proof of Theorem 2

Proof. Inspired by [4], we replace the one-hot vector in DPO with Plackett-Luce model [7], and then the online DPO can be expressed as

$$\mathcal{L}_{\text{DPO-On}}(\theta) = -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}\sim\bar{\pi}_\theta(\cdot|c)}\Big[\sum_{k=1}^{K}p_r(x_k|c)\log\frac{(\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)})^\beta}{\sum_{j=k}^{K}(\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)})^\beta}\Big], \tag{S8}$$

where $p_r(x^k_{\bar{\pi}_\theta}|c) = \frac{\exp r(x^k_{\bar{\pi}_\theta},c)}{\sum_{i=k}^K \exp r(x^i_{\bar{\pi}_\theta},c)}$.Then we have:

$$\begin{aligned}
\mathcal{L}_{\text{on}}(\theta) &= -\mathbb{E}_{c\sim\mathcal{D},x^{1:K}\sim\bar{\pi}_\theta(\cdot|c)}\Big[\sum_{k=1}^K p_r(x_k|c)\log\frac{\big(\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)}\big)^\beta}{\sum_{j=k}^K\big(\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)}\big)^\beta}\Big]\\
&= -\mathbb{E}_{c\sim\mathcal{D}}\ p_{\bar{\pi}_\theta}(x^{1:K}|c)\Big[\sum_{k=1}^K p_r(x_k|c)\log\frac{\big(\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)}\big)^\beta}{\sum_{j=k}^K\big(\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)}\big)^\beta}\Big]\\
&= -\mathbb{E}_{c\sim\mathcal{D}}\ p_{\bar{\pi}_\theta}(x^{1:K}|c)\Big[\sum_{k=1}^K p_r(x^k|c)\log\underbrace{\frac{\big(\frac{\pi_\theta(x^k|c)}{\pi_{\text{ref}}(x^k|c)}\big)^\beta}{\sum_{j=k}^K\big(\frac{\pi_\theta(x^j|c)}{\pi_{\text{ref}}(x^j|c)}\big)^\beta}}_{p_\theta(x^k|c)}\Big]\\
&= -\mathbb{E}_{c\sim\mathcal{D}}\ p_{\bar{\pi}_\theta}(x^{1:K}|c)\Big[\sum_{k=1}^K p_r(x^k|c)\log p_\theta(x^k|c)\Big]\\
&= -\mathbb{E}_{c\sim\mathcal{D}}\ p_{\bar{\pi}_\theta}(x^{1:K}|c)\Big[\sum_{k=1}^K p_r(x^k|c)(\log p_\theta(x^k|c)-\log p_r(x^k|c)+\log p_r(x^k|c))\Big]\\
&= \mathbb{E}_{c\sim\mathcal{D}}\ p_{\bar{\pi}_\theta}(x^{1:K}|c)\Big[D_{KL}(p_r|p_\theta)-p_r(x^k|c)\log p_r(x^k|c)\Big]
\end{aligned}\tag{S9}$$

Therefore, we have:

$$\nabla_\theta\mathcal{L}_{\text{on}}(\theta) = \mathbb{E}_{c\sim\mathcal{D}}\nabla_\theta\ p_{\bar{\pi}_\theta}(x^{1:K}|c)D_{KL}(p_r||p_\theta).\tag{S10}$$

The proof is completed. □

Given a sample x with a tiny generative probability $p_{\bar{\pi}_\theta|c}(x) \to 0$, and large reward value $r(x,c) \to 1$, we have $\lim_{p_{\pi_\theta}(x|c)\to 0, r(x,c)\to 1}\nabla_\theta\mathcal{L}_{\text{on}} = \mathbf{0}$.

Proof. Since x is contained in the sampled motion group $x^{1:K}$, we have:

$$\begin{aligned}
&\lim_{p_{\pi_\theta}(x|c)\to 0, r(x,c)\to 1}\nabla_\theta\mathcal{L}_{\text{on}}\\
=&\lim_{p_{\pi_\theta}(x|c)\to 0, r(x,c)\to 1}\nabla_\theta\ p_{\bar{\pi}_\theta}(x^{1:K}|c)D_{KL}(p_r||p_\theta)\\
\overset{①}{=}&\lim_{p_{\pi_\theta}(x^{1:K}|c)\to 0, r(x,c)\to 1}\nabla_\theta\ p_{\bar{\pi}_\theta}(x^{1:K}|c)D_{KL}(p_r||p_\theta)\\
=&\mathbf{0},
\end{aligned}\tag{S11}$$

where ① holds since $p_{\pi_\theta}(x^{1:K}|c) = p_{\pi_\theta}(x|c)p_{\pi_\theta}(x^M|c) \leq p_{\pi_\theta}(x|c)$, and x^M denotes a motion group obtained by removing the given motion x from the group $x^{1:K}$, i.e., satisfying that $x^M = x^{1:K} - \{x\}$. The proof is completed. □

B.3. Proof of DSoPo

Proof. Eq. (10) suggests that DSoPo samples multiple unpreferred motion candidates instead of a single unpreferred motion. Thus, we should first extend Eq. (9) as:

$$\mathcal{L}_{\text{DSoPo}}(\theta) = -\ \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta(x|c)}\log\sigma\Big(\beta\mathcal{H}_\theta(x^w,x^l,c)\Big),\tag{S12}$$

where $x^l = \mathrm{argmin}_{\{x^k_{\bar{\pi}_\theta}\}_{k=1}^K \sim \pi_\theta} r(x^k_{\pi_\theta}, c)$. Then, we have:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{DSoPo}}(\theta) =& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta(x|c)} \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}} \underbrace{p_{\bar{\pi}_\theta}(x^{1:K}|c)}_{\text{Substituting with (11)}} \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}} \Big(p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)p_\tau(r(x^l,c){\geq}\tau) + p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)p_\tau(r(x^l,c){<}\tau)\Big) \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}} \underbrace{p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)p_\tau(r(x^l,c){\geq}\tau)}_{p^{hu}_{\bar{\pi}_\theta}(x^{1:K}|c)} \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}} \underbrace{p_{\bar{\pi}_\theta}(x^{1:K}_{\bar{\pi}_\theta}|c)p_\tau(r(x^l,c){<}\tau)}_{p^{vu}_{\bar{\pi}_\theta}(x^{1:K}|c)} \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}} Z_{hu}(c)p^{hu}_{\bar{\pi}_\theta}(x^{1:K}|c) \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}\mathbb{E}_{x^{1:K}} Z_{vu}(c)p^{vu*}_{\bar{\pi}_\theta}(x^{1:K}|c) \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{hu}(c)\mathbb{E}_{x^{1:K}} p^{hu*}_{\bar{\pi}_\theta}(x^{1:K}|c) \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{vu}(c)\mathbb{E}_{x^{1:K}} p^{vu*}_{\bar{\pi}_\theta}(x^{1:K}|c) \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{hu}(c)\mathbb{E}_{x^{1:K}\sim\bar{\pi}^{hu*}_\theta} \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
& - \mathbb{E}_{(x^w,c)\sim\mathcal{D}}Z_{vu}(c)\mathbb{E}_{x^{1:K}\sim\bar{\pi}^{vu*}_\theta} \log\sigma\Big(\beta\mathcal{H}_\theta(x^w, x^l, c)\Big) \\
=& \mathcal{L}_{\mathrm{vu}}(\theta) + \mathcal{L}_{\mathrm{hu}}(\theta),
\end{aligned}
\tag{S13}
$$

where $p_{\bar{\pi}^{vu*}_\theta}(\cdot) = \frac{p^{vu}_{\bar{\pi}_\theta}(\cdot)}{Z_{vu}(c)}$ 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. The proof is completed. □

Accordingly, we rewrite $\mathcal{L}_{\mathrm{hu}}(\theta)$ and obtain the objective function of USoPo:

$$
\begin{aligned}
\mathcal{L}_{\mathrm{USoPo-hu}}(\theta) &= -\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \log\sigma\Big(\beta h_\theta(x^w, c)\Big), \\
\mathcal{L}_{\mathrm{USoPo}}(\theta) &= \mathcal{L}_{\mathrm{USoPo-hu}}(\theta) + \mathcal{L}_{\mathrm{vu}}(\theta).
\end{aligned}
\tag{S14}
$$

Implementation Now, we discuss how to deal with the computation of $Z_{vu}(c)$ and $Z_{hu}(c)$ in our implementation. As discussed in Sec. A, directly optimizing the objective function $\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo}}(\theta)$ is challenging, and we used **Algorithm 1** optimized its equivalent form:

$$
\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo}}(\theta) = -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D},x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}_\theta(\cdot|c)}
\begin{cases}
\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\mathrm{ref},x^w_t) - \beta\mathcal{L}(\theta,\mathrm{ref},x^l_t))\big)\Big), & \text{If } r(x^l,c) < \tau, \\
\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\mathrm{ref},x^w_t)\Big), & \text{Otherwise.}
\end{cases}
\tag{S15}
$$

Similarly, we can optimize the equivalent form of UDoPo to avoid the computation of $Z_{vu}(c)$ and $Z_{hu}(c)$:

$$
\mathcal{L}_{\mathrm{USoPo}}(\theta) = -\mathbb{E}_{(x^w,c)\sim\mathcal{D},x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}_\theta(\cdot|c)}
\begin{cases}
\log\sigma\Big(\beta\mathcal{H}_\theta(x^w,x^l,c)\Big), & \text{If } r(x^l,c) < \tau, \\
\log\sigma\Big(\beta h_\theta(x^w,c)\Big), & \text{Otherwise.}
\end{cases}
\tag{S16}
$$

The proof of Eq. (S16) follows the same steps as the proof of Eq. (S15) in Sec. A.

B.4. Discussion of USoPo and DSoPo

In this section, we discuss the relationship between USoPo and DSoPo and the difference between their optimization. Here, USoPo and DSoPo are defined as:

$$\mathcal{L}_{\text{USoPo}}(\theta) = -\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \log \sigma\Big(\beta h_\theta(x^w, c)\Big) + \mathcal{L}_{\text{vu}}(\theta). \tag{S17}$$

$$\mathcal{L}_{\text{DSoPo}}(\theta) = \mathcal{L}_{\text{vu}}(\theta) + \mathcal{L}_{\text{hu}}(\theta), \tag{S18}$$

Relationship between USoPo and DSoPo We begin by analyzing the size relationship between USoPo and DSoPo:

$$\begin{aligned}
&\mathcal{L}_{\text{DSoPo}}(\theta) - \mathcal{L}_{\text{USoPo}}(\theta)\\
=&\mathcal{L}_{\text{hu}}(\theta) + \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \log \sigma\Big(\beta h_\theta(x^w, c)\Big)\\
=&-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \log \sigma\Big(\beta \mathcal{H}_\theta(x^w, x^l, c)\Big) + \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \log \sigma\Big(\beta h_\theta(x^w, c)\Big)\\
=&-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \Big[\log \sigma\Big(\beta \mathcal{H}_\theta(x^w, x^l, c)\Big) - \log \sigma\Big(\beta h_\theta(x^w, c)\Big)\Big].
\end{aligned} \tag{S19}$$

Considering that $\mathcal{H}_\theta(x^w, x^l, c) = h_\theta(x^w, c) - h_\theta(x^l, c)$ and $h_\theta(x, c) = \log \frac{\pi_\theta(x|c)}{\pi_{\text{ref}}(x|c)}$, we have:

$$\begin{aligned}
&\mathcal{L}_{\text{DSoPo}}(\theta) - \mathcal{L}_{\text{USoPo}}(\theta)\\
=&-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \Big[\log \sigma\Big(\beta \mathcal{H}_\theta(x^w, x^l, c)\Big) - \log \sigma\Big(\beta h_\theta(x^w, c)\Big)\Big]\\
=&-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \Big[\log \frac{\exp \beta h_\theta(x^w, c)}{\exp \beta h_\theta(x^w, c) + \exp \beta h_\theta(x^l, c)} - \log \frac{\exp \beta h_\theta(x^w, c)}{\exp \beta h_\theta(x^w, c) + 1}\Big]\\
=&-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \Big[\log \frac{\exp \beta h_\theta(x^w, c) + 1}{\exp \beta h_\theta(x^w, c) + \exp \beta h_\theta(x^l, c)}\Big]\\
=&-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \Big[\log \frac{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta}\Big].
\end{aligned} \tag{S20}$$

In general, DPO focuses on reducing the generative probability of loss samples (unpreferred motions). Consequently, the generative probability of the policy model $\pi_\theta(x^l|c)$ will be lower than that of the reference model $\pi_{\text{ref}}(x^l|c)$, i.e., $\pi_\theta(x^l|c) \leq \pi_{\text{ref}}(x^l|c)$, resulting in $\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)} \leq 1$. Hence, the following relationship holds:

$$\begin{aligned}
&\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)} \leq 1\\
\Rightarrow&(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1 \geq (\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta\\
\Rightarrow&\frac{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta} \geq 1\\
\Rightarrow&\log \frac{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta} \geq 0\\
\Rightarrow&\underbrace{-\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_\theta^{hu*}} \Big[\log \frac{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta}\Big]}_{\mathcal{L}_{\text{DSoPo}}(\theta) - \mathcal{L}_{\text{USoPo}}(\theta)} \leq 0\\
\Rightarrow&\mathcal{L}_{\text{DSoPo}}(\theta) \leq \mathcal{L}_{\text{USoPo}}(\theta).
\end{aligned} \tag{S21}$$

Eq. (S21) indicates that $\mathcal{L}_{\text{USoPo}}$ is one of upper bounds of $\mathcal{L}_{\text{DSoPo}}$.

Difference between the optimization of USoPo and DSoPo The difference between the optimization of USoPo and DSoPo can be measured by that between their objective function. Let $\mathcal{L}_{\text{d}}(\theta) = \mathcal{L}_{\text{USoPo}}(\theta) - \mathcal{L}_{\text{DSoPo}}(\theta)$, the difference between their objective function can be denoted as:

$$\begin{aligned}\mathcal{L}_{\text{d}}(\theta) =& \mathcal{L}_{\text{USoPo}}(\theta) - \mathcal{L}_{\text{DSoPo}}(\theta)\\ =& \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_{\theta}^{hu*}} \left[\log \frac{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta}\right] \overset{\textcircled{1}}{\geq} 0\end{aligned} \tag{S22}$$

where ① holds due to Eq. (S21). As discussed above, the generative probability of the policy model $\pi_\theta(x^l|c)$ will be lower than that of the reference model $\pi_{\text{ref}}(x^l|c)$, and thus $\pi_\theta(x^l|c)$ falls in the range between 0 and $\pi_{\text{ref}}(x^l|c)$, i.e., $0 \leq \pi_\theta(x^l|c) \leq \pi_{\text{ref}}(x^l|c)$.

Assuming that the value of $\pi_\theta(x^w|c)$ is fixed, the value of $\mathcal{L}_{\text{d}}(\theta)$ is negatively correlated with $\pi_\theta(x^l|c)$, since we have:

$$\begin{aligned}\nabla_\theta \mathcal{L}_{\text{d}}(\theta) =& \nabla_\theta - \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_{\theta}^{hu*}} \left[\log \frac{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + 1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta}\right]\\ =& \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_{\theta}^{hu*}} \nabla_\theta - \log\left[(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta\right]\\ =& \mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c) \mathbb{E}_{x^{1:K}\sim\bar{\pi}_{\theta}^{hu*}} \frac{1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta} - \nabla_\theta(\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta\\ \overset{\textcircled{1}}{\sim}& - \nabla_\theta(\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta.\end{aligned} \tag{S23}$$

where ① holds since $\frac{1}{(\frac{\pi_\theta(x^w|c)}{\pi_{\text{ref}}(x^w|c)})^\beta + (\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)})^\beta} > 0$.

Hence, when the generative probability of unpreferred motions $\pi_\theta(x^l|c)$ is lower, the difference between the optimization of USoPo and DSoPo is larger. However, the unpreferred motions are sampled from the relatively high-preference distribution $\pi_{\bar{\theta}}^{hu*}$, and thus should not be treated as unpreferred motions. Using $\mathcal{L}_{\text{USoPo}}(\theta)$ to optimize policy model π_θ instead of $\mathcal{L}_{\text{DSoPo}}(\theta)$ can avoid unnecessary optimization of these relatively high-preference unpreferred motion $\mathcal{L}_{\text{d}}(\theta)$.

B.5. Proof of Eq. (16)

Before proving Eq. (16), we first present some useful lemmas from [10].

Lemma 1. *[10] Given a winning sample x_w and a losing sample x_l, the DPO denoted as*

$$\mathcal{L}_{\text{DPO}}(\theta) = \mathbb{E}_{(x^w,x^l,c)\sim\mathcal{D}}\left[-\log\sigma\left(\beta\log\frac{\pi_\theta(x^w|c)}{\pi_{ref}(x^w|c)} - \beta\log\frac{\pi_\theta(x^l|c)}{\pi_{\text{ref}}(x^l|c)}\right)\right]. \tag{S24}$$

Then the objective function for diffusion models can be denoted as:

$$\mathcal{L}_{\textit{DPO-Diffusion}}(\theta) = -\mathbb{E}_{(x_0^w,x_0^l)\sim\mathcal{D}} \log\sigma(\beta\mathbb{E}_{x_{1:T}^w\sim\pi_\theta(x_{1:T}^w|x_0^w),x_{1:T}^l\sim\pi_\theta(x_{1:T}^l|x_0^l)}[\log\frac{\pi_\theta(x_{0:T}^w)}{\pi_{\text{ref}}(x_{0:T}^w)} - \log\frac{\pi_\theta(x_{0:T}^l)}{\pi_{\text{ref}}(x_{0:T}^l)}]), \tag{S25}$$

where x_t^* denoted the noised sample x^* for the t-th step.

Lemma 2. *[10] Given the objective function of diffusion-based DPO denoted as Eq. (S25), it has an upper bound $\mathcal{L}_{\text{UB}}(\theta)$:*

$$\begin{aligned}\mathcal{L}_{\textit{DPO-Diffusion}}(\theta) \leq& - \mathbb{E}_{(x_0^w,x_0^l)\sim\mathcal{D},t\sim\mathcal{U}(0,T),x_{t-1,t}^w\sim\pi_\theta(x_{t-1,t}^w|x_0^w),x_{t-1,t}^l\sim\pi_\theta(x_{t-1,t}^l|x_0^l)} \log\sigma\\ & \underbrace{\left(\beta T\log\frac{\pi_\theta(x_{t-1}^w|x_t^w)}{\pi_{\text{ref}}(x_{t-1}^w|x_t^w)} - \beta T\log\frac{\pi_\theta(x_{t-1}^l|x_t^l)}{\pi_{\text{ref}}(x_{t-1}^l|x_t^l)}\right),}_{\mathcal{L}_{\text{UB}}(\theta)}\end{aligned} \tag{S26}$$

where T denotes the number of diffusion steps.

Lemma 3. *[10] Given the objective function for diffusion model denoted as Eq. (S26), it can be rewritten as :*

$$\begin{aligned}\mathcal{L}_{\mathrm{UB}}(\theta) = -\mathbb{E}_{(x_0^w, x_0^l)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), x_t^w\sim q(x_t^w|x_0^w), x_t^l\sim q(x_t^l|x_0^l)} \log\sigma(-\beta T\omega_t \\ (\|\epsilon-\epsilon_\theta(x_t^w,t)\|_2^2 - \|\epsilon-\epsilon_{\mathrm{ref}}(x_t^w,t)\|_2^2 - \big(\|\epsilon-\epsilon_\theta(x_t^l,t)\|_2^2 - \|\epsilon-\epsilon_{\mathrm{ref}}(x_t^l,t)\|_2^2)\big)),\end{aligned} \tag{S27}$$

where $x_t^* = \alpha_t x_0^* + \sigma_t\epsilon$, $\epsilon\sim\mathcal{N}(0,\mathbb{I})$ is a draw from the distribution of forward process $q(x_t^*|x_0^*)$.

Now, we proof Eq. (16) based on these lemmas.

Proof. This proof has three steps. In each step, we apply the three lemmas introduced above in succession. We begin with the loss function of SoPo for probability models:

$$\begin{aligned}\mathcal{L}_{\mathrm{SoPo}}(\theta) = & \underbrace{-\mathbb{E}_{(x^w,c)\sim\mathcal{D}, x^{1:K}_{\bar\pi_\theta}\sim\bar\pi_\theta^{vu*}(\cdot|c)} Z_{vu}(c)\Big[\log\sigma\Big(\beta_w(x^w)h_\theta(x^w,c) - \beta h_\theta(x^l,c)\Big)\Big]}_{\mathcal{L}_{\mathrm{SoPo-vu}}(\theta)} \\ & -\underbrace{\mathbb{E}_{(x^w,c)\sim\mathcal{D}} Z_{hu}(c)\log\sigma\Big(\beta_w(x^w)h_\theta(x^w,c)\Big)}_{\mathcal{L}_{\mathrm{SoPo-hu}}(\theta)}.\end{aligned} \tag{S28}$$

Based on **Lemma 1**, we can rewrite the objective function for diffusion models:

$$\begin{aligned}\mathcal{L}_{\mathrm{SoPo-Diffusion}}(\theta) &= \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-vu}}(\theta) + \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-hu}}(\theta) \\ \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-vu}}(\theta) &= -\mathbb{E}_{(x_0^w,c)\sim\mathcal{D}, x_0^{1:K}\sim\bar\pi_\theta^{vu*}(\cdot|c)} Z_{vu}(c) \\ & \log\sigma(\mathbb{E}_{x_{1:T}^w\sim\pi_\theta(x_{1:T}^w|x_0^w), x_{1:T}^l\sim\pi_\theta(x_{1:T}^l|x_0^l)}[\beta_w(x_0^w)\log\frac{\pi_\theta(x_{0:T}^w)}{\pi_{\mathrm{ref}}(x_{0:T}^w)} - \beta\log\frac{\pi_\theta(x_{0:T}^l)}{\pi_{\mathrm{ref}}(x_{0:T}^l)}]), \\ \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-hu}}(\theta) &= -\mathbb{E}_{(x_0^w,c)\sim\mathcal{D}} Z_{hu}(c)\log\sigma(\mathbb{E}_{x_{1:T}^w\sim\pi_\theta(x_{1:T}^w|x_0^w)}[\beta_w(x^w)\log\frac{\pi_\theta(x_{0:T}^w)}{\pi_{\mathrm{ref}}(x_{0:T}^w)}]),\end{aligned} \tag{S29}$$

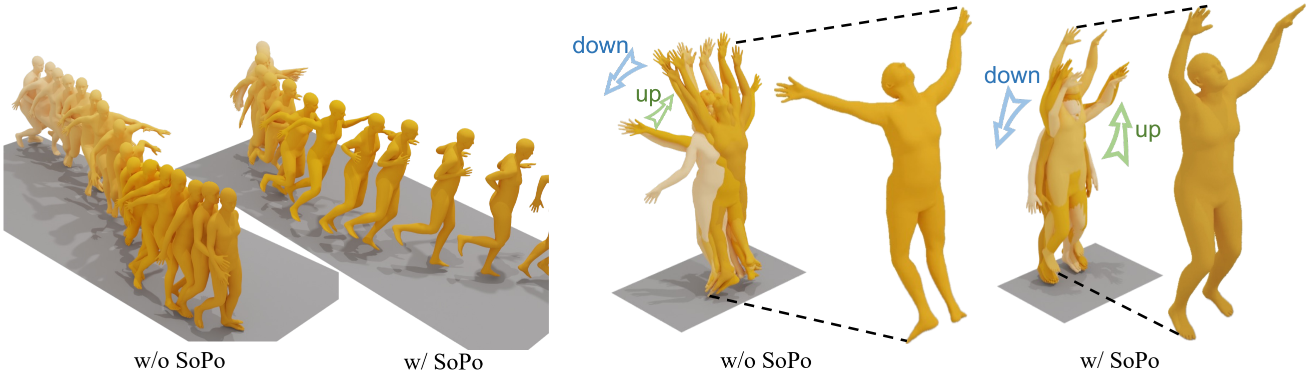
where x_t^* denoted the noised sample x^* for the t-th step. According to **Lemma 2**, the upper bound of $\mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-vu}}(\theta)$ and $\mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-hu}}(\theta)$ can be denoted as:

$$\begin{aligned}\mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-vu}}(\theta) \leq & \underbrace{-\mathbb{E}_{(x_0^w,c)\sim\mathcal{D}, x_0^{1:K}\sim\bar\pi_\theta^{vu*}(\cdot|c), t\sim\mathcal{U}(0,T), x_{t-1,t}^w\sim\pi_\theta(x_{t-1,t}^w|x_0^w), x_{t-1,t}^l\sim\pi_\theta(x_{t-1,t}^l|x_0^l)} \log\sigma\left(\beta_w(x_0^w)T\log\frac{\pi_\theta(x_{t-1}^w|x_t^w)}{\pi_{\mathrm{ref}}(x_{t-1}^w|x_t^w)} - \beta T\log\frac{\pi_\theta(x_{t-1}^l|x_t^l)}{\pi_{\mathrm{ref}}(x_{t-1}^l|x_t^l)}\right)}_{\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}}(\theta)}, \\ \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-hu}}(\theta) \leq & \underbrace{-\mathbb{E}_{(x_0^w,c)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), x_{t-1,t}^w\sim\pi_\theta(x_{t-1,t}^w|x_0^w)} \log\sigma\left(\beta_w(x_0^w)T\log\frac{\pi_\theta(x_{t-1}^w|x_t^w)}{\pi_{\mathrm{ref}}(x_{t-1}^w|x_t^w)}\right)}_{\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}}(\theta)}, \\ \mathcal{L}_{\mathrm{SoPo-Diffusion}}(\theta) = & \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-vu}}(\theta) + \mathcal{L}^{\mathrm{diff-ori}}_{\mathrm{SoPo-hu}}(\theta) \leq \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}}(\theta) + \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}}(\theta) = \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo}}(\theta).\end{aligned} \tag{S30}$$

Applying **Lemma 3** to $\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}}(\theta)$ and $\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}}(\theta)$, we have

$$\begin{aligned}\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}}(\theta) = & -\mathbb{E}_{(x_0^w,c)\sim\mathcal{D}, x_0^{1:K}\sim\bar\pi_\theta^{vu*}(\cdot|c), t\sim\mathcal{U}(0,T), x_t^w\sim q(x_t^w|x_0^w), x_t^l\sim q(x_t^l|x_0^l)}\log\sigma\Bigg(-T\omega_t \\ & \Big(\beta_w(x_0^w)\big(\|\epsilon-\epsilon_\theta(x_t^w,t)\|_2^2 - \|\epsilon-\epsilon_{\mathrm{ref}}(x_t^w,t)\|_2^2\big) - \beta\big(\|\epsilon-\epsilon_\theta(x_t^l,t)\|_2^2 - \|\epsilon-\epsilon_{\mathrm{ref}}(x_t^l,t)\|_2^2\big)\Big)\Bigg), \\ \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}}(\theta) = & -\mathbb{E}_{(x_0^w,c)\sim\mathcal{D}, t\sim\mathcal{U}(0,T), x_{t-1,t}^w\sim\pi_\theta(x_{t-1,t}^w|x_0^w)}\log\sigma\left(-T\omega_t\beta_w(x_0^w)(\|\epsilon-\epsilon_\theta(x_t^w,t)\|_2^2 - \|\epsilon-\epsilon_{\mathrm{ref}}(x_t^w,t)\|_2^2)\right)\end{aligned} \tag{S31}$$

$$\mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo}}(\theta) = \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-vu}}(\theta) + \mathcal{L}^{\mathrm{diff}}_{\mathrm{SoPo-hu}}(\theta) \tag{S32}$$



(a) A person runs to their right and then curves to the left and continues to run then stops.

(b) A man jumps and brings both arms above his head as … and then moves them back into the original position.

Figure S2. Visual results on HumanML3D dataset.

To simplify the symbolism, the objective functions can be rewritten as:

$$\begin{aligned}\mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{vu}} &= -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D},x^{1:K}_{\bar{\pi}_\theta}\sim\bar{\pi}^{vu*}_\theta}(\cdot|c)Z_{vu}(c)\Big[\log\sigma\Big(-T\omega_t\big(\beta_w(x_w)(\mathcal{L}(\theta,\text{ref},x^w_t)-\beta\mathcal{L}(\theta,\text{ref},x^l_t))\Big)\Big]\\ \mathcal{L}^{\text{diff}}_{\text{SoPo}-\text{hu}} &= -\mathbb{E}_{t\sim\mathcal{U}(0,T),(x^w,c)\sim\mathcal{D}}Z_{hu}(c)\Big[\log\sigma\Big(-T\omega_t\beta_w(x_w)\mathcal{L}(\theta,\text{ref},x^w_t)\Big)\Big]\end{aligned} \tag{S33}$$

where $\mathcal{L}(\theta,\text{ref},x_t)=\mathcal{L}(\theta,x_t)-\mathcal{L}(\text{ref},x_t)$, and $\mathcal{L}(\theta/\text{ref},x_t)=\|\epsilon_{\theta/\text{ref}}(x_t,t)-\epsilon\|_2^2$ denotes the loss of the policy or reference model. The proof is completed. □

C. Experiment

C.1. Additional Experimental Datails

Datasets & Evaluation HumanML3D is derived from the AMASS [6] and HumanAct12 [3] datasets and contains 14,616 motions, each described by three textual annotations. All motion is split into train, test, and evaluate sets, composed of 23384, 1460, and 4380 motions, respectively. For both HumanML3D and KIT-ML datasets, we follow the official split and report the evaluated performance on the test set.

We evaluate our experimental results on two main aspects: alignment quality and generation quality. Following prior research [2, 8, 11], we use motion retrieval precision (R-Precision) and multi-modal distance (MM Dist) to evaluate alignment quality, while diversity and Fréchet Inception Distance (FID) are employed to assess generation quality. (1) R-Precision evaluates the similarity between generated motion and their corresponding text descriptions. Higher values indicate better alignment quality. (2) MM Dist represents the average distance between the generated motion features and their corresponding text embedding. (3) Diversity calculates the variation in generated samples. A diversity close to real motions ensures that the model produces rich patterns rather than repetitive motions. (4) FID measures the distribution proximity between the generated and real samples in latent space. Lower FID scores indicate higher generation quality.

Implementation Details For the preference alignment of MDM [9], we largely adopt the original implementation's settings. The model is trained using the AdamW optimizer [5] with a cosine decay learning rate scheduler and linear warm-up over the initial steps. We use a batch size of 64 and a learning rate of 10^{-5}, with a guidance parameter of 2.5 during testing. Diffusion employs a cosine noise schedule with 50 steps, and an evaluation batch size of 32 ensures consistent metric computation. For fine-tuning MLD [1], we similarly follow its original parameter settings.

C.2. Additional Experimental Results

Visualization We visualize the generated motion for our SoPo. As shown in Fig. S2, our proposed approach helps text-to-motion models avoid frequent mistakes, such as incorrect movement direction and specific semantics. More results can be found in Appendix C.2. Additionally, we also present additional results generated by text-to-motion models with SoPo, as illustrated in Fig. S1. Our proposed SoPo significantly enhances the ability of text-to-motion models to comprehend text semantics. For instance, in Fig. S1 (j), a model integrated with SoPo can successfully interpret the semantics of "zig-zag pattern", whereas a model without SoPo struggles to do so.

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, 10

[2] 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. 10

[3] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the ACM International Conference on Multimedia*, page 2021–2029, New York, NY, USA, 2020. Association for Computing Machinery. 10

[4] Cheng Lu Haozhe Ji, Pei Ke Yilin Niu, Jun Zhu Hongning Wang, and Minlie Huang Jie Tang. Towards efficient exact optimization of language model alignment. *The Forty-first International Conference on Machine Learning*, 2024. 4

[5] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 10

[6] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael Black. AMASS: Archive of Motion Capture As Surface Shapes . In *IEEE/CVF International Conference on Computer Vision*, pages 5441–5450, Los Alamitos, CA, USA, 2019. IEEE Computer Society. 10

[7] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24(2):193–202, 1975. 3, 4

[8] Zeping Ren, Shaoli Huang, and Xiu Li. Realistic human motion generation with cross-diffusion models. *European Conference on Computer Vision*, 2024. 10

[9] 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. 10

[10] 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. 8, 9

[11] 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. 10