Human Joint Kinematics Diffusion-Refinement for Stochastic Motion Prediction ——Supplementary Material——

Dong Wei[1], Huaijiang Sun[1][*], Bin Li[2], Jianfeng Lu[1], Weiqing Li[1], Xiaoning Sun[1], Shengxiang Hu[1]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[2] Tianjin AiForward Science and Technology Co., Ltd., Tianjin, China
{csdwei, sunhuaijiang, lujf, li_weiqing, sunxiaoning, hushengxiang}@njust.edu.cn, libin@aiforward.com

Detailed Derivations

Properties of Diffusion Model

Suppose that $\beta_1, \beta_2, \cdots, \beta_K$ are pre-determined increasing variance schedulers; $\alpha_k = 1 - \beta_k$ and $\bar{\alpha}_k = \prod_{s=1}^{k} \alpha_s$. The following two properties are necessary to derive the final training objective $L(\psi, \theta)$.

Property 1. The marginal of the forward diffusion process is tractable, and can be derived as:

$$\begin{aligned} q(\mathcal{X}^k|\mathcal{X}^0) &= \int q(\mathcal{X}^{1:k}|\mathcal{X}^0) d\mathcal{X}^{1:(k-1)} \\ &= \mathcal{N}(\mathcal{X}^k; \sqrt{\bar{\alpha}_k}\mathcal{X}^0, (1-\bar{\alpha}_k)\mathbf{I}). \end{aligned} \tag{1}$$

This property is proved in the supplementary material of (Ho, Jain, and Abbeel 2020), and therefore we can derive the following closed-form solution of $\mathcal{X}^k$ given $\mathcal{X}^0$:

$$\mathcal{X}^k = \sqrt{\bar{\alpha}_k}\mathcal{X}^0 + \sqrt{1-\bar{\alpha}_k}\epsilon, \tag{2}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a random variable.

Property 2. The posterior of the forward diffusion process is tractable, and can be derived by Bayes' rule as follows:

$$q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0) = \frac{q(\mathcal{X}^k|\mathcal{X}^{k-1}, \mathcal{X}^0) q(\mathcal{X}^{k-1}|\mathcal{X}^0)}{q(\mathcal{X}^k|\mathcal{X}^0)}. \tag{3}$$

Since the three probabilities on the right are Gaussian, the obtained posterior is also Gaussian, which is formulated as:

$$q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0) = \mathcal{N}(\mathcal{X}^{k-1}; \tilde{\boldsymbol{\mu}}_k(\mathcal{X}^k, \mathcal{X}^0), \tilde{\beta}_k\mathbf{I}), \tag{4}$$

where

$$\tilde{\boldsymbol{\mu}}_k(\mathcal{X}^k, \mathcal{X}^0) := \frac{\sqrt{\bar{\alpha}_{k-1}}\beta_k}{1-\bar{\alpha}_k}\mathcal{X}^0 + \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k}\mathcal{X}^k, \tag{5}$$

and

$$\tilde{\beta}_k := \frac{1-\bar{\alpha}_{k-1}}{1-\bar{\alpha}_k}\beta_k. \tag{6}$$

Combined with (2) and (5), we can reformulate the mean of the posterior by using $\mathcal{X}^k$ and ϵ as follows:

$$\begin{aligned} \tilde{\boldsymbol{\mu}}_k(\mathcal{X}^k(\mathcal{X}^0, \epsilon), \epsilon) &= (\frac{\sqrt{\bar{\alpha}_{k-1}}\beta_k}{\sqrt{\bar{\alpha}_k}(1-\bar{\alpha}_k)} + \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k})\mathcal{X}^k(\mathcal{X}^0, \epsilon) \\ &+ \frac{\sqrt{1-\bar{\alpha}_k}\sqrt{\bar{\alpha}_{k-1}}\beta_k}{\sqrt{\bar{\alpha}_k}(1-\bar{\alpha}_k)}\epsilon \\ &= \frac{\sqrt{\bar{\alpha}_{k-1}}(\beta_k + \alpha_k(1-\bar{\alpha}_{k-1}))}{\sqrt{\bar{\alpha}_k}(1-\bar{\alpha}_k)}\mathcal{X}^k(\mathcal{X}^0, \epsilon) \\ &- \frac{\beta_k}{\sqrt{\alpha_k}\sqrt{1-\bar{\alpha}_k}}\epsilon \\ &= \frac{1}{\sqrt{\alpha_k}}(\mathcal{X}^k(\mathcal{X}^0, \epsilon) - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon). \end{aligned} \tag{7}$$

Derivations of $L(\psi, \theta)$

$$\begin{aligned} L(\psi, \theta) =& \mathbb{E}_q \left[-\log p(\mathcal{X}^K) - \sum_{k=1}^{K} \log \frac{p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{C})}{q(\mathcal{X}^k|\mathcal{X}^{k-1})} \right] \\ =& \mathbb{E}_q \left[-\log p(\mathcal{X}^K) - \log \frac{p_\theta(\mathcal{X}^0|\mathcal{X}^1, \mathcal{C})}{q(\mathcal{X}^1|\mathcal{X}^0)} \right. \\ & \left. - \sum_{k=2}^{K} \log \frac{p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{C})}{q(\mathcal{X}^k|\mathcal{X}^{k-1})} \right] \\ =& \mathbb{E}_q \left[-\log p(\mathcal{X}^K) - \log \frac{p_\theta(\mathcal{X}^0|\mathcal{X}^1, \mathcal{C})}{q(\mathcal{X}^1|\mathcal{X}^0)} \right. \\ & \left. - \sum_{k=2}^{K} \log p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{C}) \frac{q(\mathcal{X}^{k-1}|\mathcal{X}^0)}{q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0) q(\mathcal{X}^k|\mathcal{X}^0)} \right] \\ =& \mathbb{E}_q \left[-\log p(\mathcal{X}^K) - \log \frac{p_\theta(\mathcal{X}^0|\mathcal{X}^1, \mathcal{C})}{q(\mathcal{X}^1|\mathcal{X}^0)} \right. \\ & \left. - \sum_{k=2}^{K} \log \frac{p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{C})}{q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0)} - \sum_{k=2}^{K} \log \frac{q(\mathcal{X}^{k-1}|\mathcal{X}^0)}{q(\mathcal{X}^k|\mathcal{X}^0)} \right] \\ =& \mathbb{E}_q \left[-\log p(\mathcal{X}^K) - \sum_{k=2}^{K} \log \frac{p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{C})}{q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0)} \right. \\ & \left. - \log p_\theta(\mathcal{X}^0|\mathcal{X}^1, \mathcal{C}) + \log q(\mathcal{X}^K|\mathcal{X}^0) \right] \\ =& \mathbb{E}_q \left[D_{KL}(q(\mathcal{X}^K|\mathcal{X}^0)||p(\mathcal{X}^K)) - \log p_\theta(\mathcal{X}^0|\mathcal{X}^1, \mathcal{C}) \right. \\ & \left. + \sum_{k=2}^{K} D_{KL}(q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0)||p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{C})) \right], \end{aligned} \tag{8}$$

[*] Corresponding author.

where $\mathcal{C}$ is the encoded feature of historical sequence $\mathcal{D}$ by using the neural network f_ψ, i.e., $\mathcal{C} = f_\psi(\mathcal{D})$.

We neglect the first term because it has no learnable parameters in diffusion process. For the last term, we need to match generative transition $p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k)$ with ground truth posterior $q(\mathcal{X}^{k-1}|\mathcal{X}^k, \mathcal{X}^0)$. As the KL divergence of two Gaussian can be regarded as the difference of the corresponding means, according to (7), we have:

$$\begin{aligned}&\mathbb{E}_{\mathcal{X}^0,\epsilon}\left[\lambda\left\|\tilde{\boldsymbol{\mu}}_k(\mathcal{X}^k(\mathcal{X}^0,\epsilon),\epsilon)-\boldsymbol{\mu}_\theta(\mathcal{X}^k,k,\mathcal{C})\right\|^2\right]\\ =&\mathbb{E}_{\mathcal{X}^0,\epsilon}\left[\lambda\left\|\frac{1}{\sqrt{\alpha_k}}(\mathcal{X}^k(\mathcal{X}^0,\epsilon)-\frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon)-\boldsymbol{\mu}_\theta(\mathcal{X}^k,k,\mathcal{C})\right\|^2\right].\end{aligned} \tag{9}$$

We also prove that the second term of (8) is tractable with the same formulation of (9) in the case of $k = 1$. More precisely, $-\log p_\theta(\mathcal{X}^0|\mathcal{X}^1, \mathcal{C})$ indicates that the outputs of prediction model should follow the distribution of real data. In other words, when $k = 1$, we should minimize the difference between the mean $\boldsymbol{\mu}_\theta(\mathcal{X}^k, k, \mathcal{C})$ and the ground truth $\mathcal{X}^0$, i.e., $\mathbb{E}[\lambda\|\mathcal{X}^0 - \boldsymbol{\mu}_\theta(\mathcal{X}^1, 1, \mathcal{C})\|^2]$. As $\bar{\alpha}_1 = \alpha_1$, we have:

$$\mathcal{X}^0 = \frac{1}{\sqrt{\bar{\alpha}_1}}(\mathcal{X}^1 - \sqrt{1-\bar{\alpha}_1}\epsilon) = \frac{1}{\sqrt{\alpha_1}}(\mathcal{X}^1 - \frac{\beta_1}{\sqrt{1-\bar{\alpha}_1}}\epsilon).$$

Therefore, (9) holds for both $k = 1$ and $k \geq 2$.

As shown in (9), given the inputs $\mathcal{X}^k(\mathcal{X}^0, \epsilon)$ and $\mathcal{C}$, $\boldsymbol{\mu}_\theta(\mathcal{X}^k, k, \mathcal{C})$ is used to predict $\frac{1}{\sqrt{\alpha_k}}(\mathcal{X}^k(\mathcal{X}^0, \epsilon) - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon)$. We instead train a neural network to predict the noise $\epsilon_\theta(\mathcal{X}^k, k, \mathcal{C})$, where

$$\boldsymbol{\mu}_\theta(\mathcal{X}^k, k, \mathcal{C}) = \frac{1}{\sqrt{\alpha_k}}(\mathcal{X}^k(\mathcal{X}^0, \epsilon) - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(\mathcal{X}^k, k, \mathcal{C})).$$

As a result, the final training objective becomes:

$$L(\psi, \theta) = \mathbb{E}_{k,\mathcal{X}^0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathcal{X}^k, k, \mathcal{C})\|^2\right],$$

Once the noise estimated network $\epsilon_\theta(\mathcal{X}^k, k, \mathcal{C})$ is obtained, the sampling process is performed by gradually sampling from $p_\theta(\mathcal{X}^{k-1}|\mathcal{X}^k)$ as $k = K, K-1, \cdots, 1$:

$$\mathcal{X}^{k-1} = \frac{1}{\sqrt{\alpha_k}}(\mathcal{X}^k - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\epsilon_\theta(\mathcal{X}^k, k, \mathcal{C})) + \sqrt{\beta_k}\mathbf{z},$$

which is similar to Langevin dynamics (Du and Mordatch 2019) in energy-based models.

Implementation Details

Training Parameters. For the diffusion network, the learning rate is set to 0.0005, and will decrease after the first 100 training epochs. For the refinement network, the learning rate is set to 0.0005 with a 0.96 decay every two epochs. The gradients are clipped to a maximum l_2-norm of 1. We define hyper-parameters β_k in the diffusion process according to a linear scheduler, i.e., $\beta_k = \frac{k-1}{K-1} \cdot (\beta_K - \beta_1)$, $k = 1, 2, \cdots, K$, where $\beta_1 = 10^{-4}$, $\beta_K = 0.05$ and $K = 100$. Compared to DDPM (Ho, Jain, and Abbeel 2020) in image generation domain with $\beta_1 = 10^{-4}$, $\beta_K = 0.02$ and $K = 1000$, the reason to increase β_k for smaller K is to make $q(\mathcal{X}^K|\mathcal{X}^0)$ close to $p(\mathcal{X}^K) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The hyper-parameters of the different terms (γ, λ_1, λ_2) are set to (0.01, 200.0, 1.0).

Reproducibility. For better understanding, we provide the code in https://github.com/csdwei/MotionDiff.

More Qualitative Results

Given an observed sequence, we show different poses of five future pose sequences generated by the proposed method in Figure 1a, and the corresponding end poses of diffusion process from K to 0 in Figure 1b. As depicted in Figure 2, our approach sometimes confuses the left and right parts of the body (e.g., when turning around, the left hand of a walking person changed over time to the right hand walking in the opposite direction). Due to the symmetry of the human body, our approach produces such deviation.

References

Du, Y.; and Mordatch, I. 2019. Implicit generation and modeling with energy based models. *Proceedings of the Advances in Neural Information Processing Systems*, 32.

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

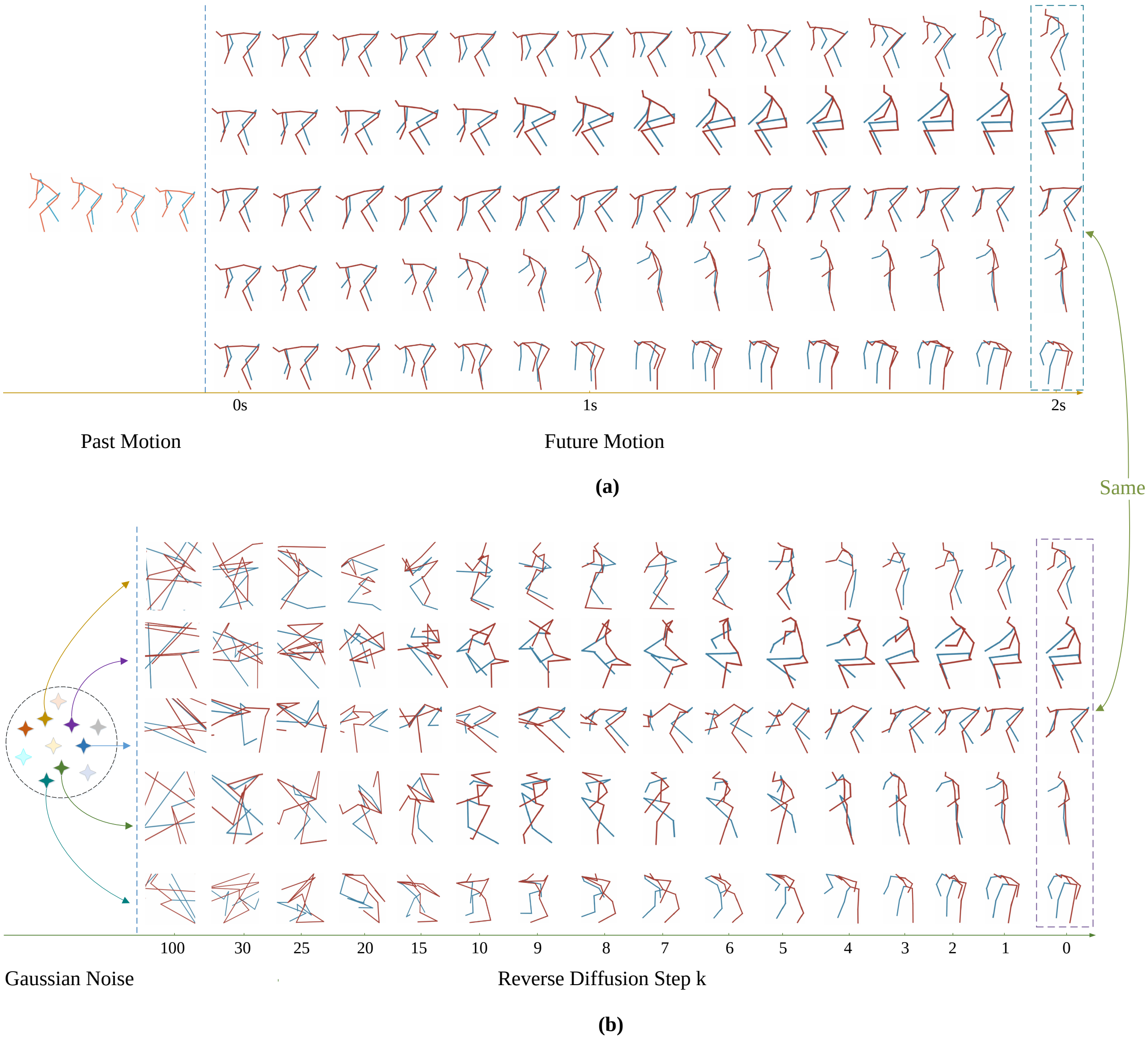

Figure 1: Visualization results of (a) stochastic human motion prediction, and (b) reverse diffusion process. Given an observed sequence, we show five possible samples of future sequences, and the corresponding reverse diffusion process of the end poses.

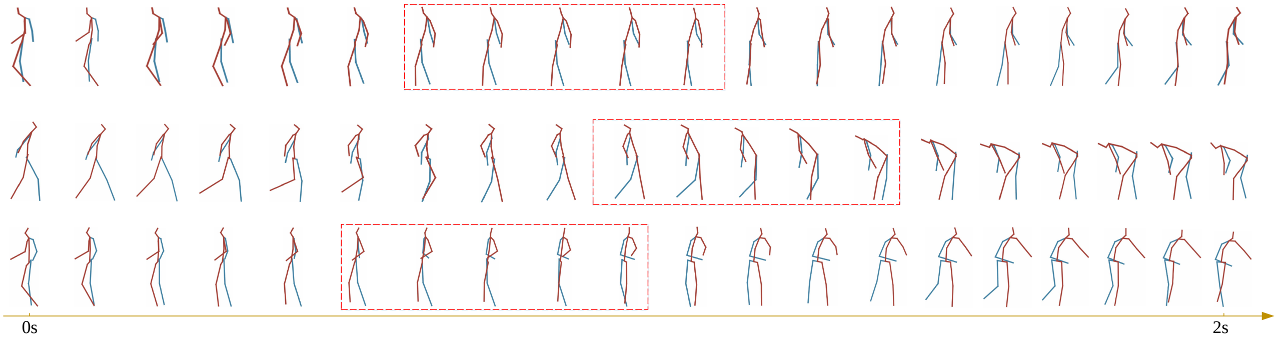

Figure 2: Qualitative results of failure cases. Our method sometimes produces unrealistic motions when a person turns around, as highlighted by red boxes.