Appendix

A. Calculation on Optimal Transport

In this section, we will provide the optimize details on optimal transport. That is, the problem definition of optimal transport is given as:

$$\min_{\boldsymbol{\pi}\in\Delta} J_{\mathrm{OT}} = \langle \boldsymbol{\pi}, \boldsymbol{C}\rangle$$
$$s.t.\ \Delta = \left\{ \sum_{j=1}^{N} \pi_{ij} = a_i, \quad \sum_{i=1}^{N} \pi_{ij} = b_j, \quad \pi_{ij} \geq 0 \right\}, \tag{14}$$

To start with, we should first figure out the Lagrange multipliers of optimal transport as:

$$\max_{\boldsymbol{f},\boldsymbol{g},\boldsymbol{s}} \min_{\boldsymbol{\pi}} \mathcal{J} = \langle \boldsymbol{f}, \boldsymbol{a}\rangle + \langle \boldsymbol{g}, \boldsymbol{b}\rangle + \left[\sum_{i,j} \left(C_{ij} - f_i - g_j - s_{ij}\right) \pi_{ij} \right] \tag{15}$$

where $\boldsymbol{f}$, $\boldsymbol{g}$ and $\boldsymbol{s}$ denote the Lagrange multipliers. By taking the differentiation on π_{ij}, we can obtain the following results as:

$$\begin{cases} \dfrac{\partial \mathcal{J}}{\partial \pi_{ij}} = C_{ij} - f_i - g_j - s_{ij} = 0 \\ s_{ij} \geq 0 \end{cases} \tag{16}$$

Note that $s_{ij} \geq 0$ and $s_{ij}\pi_{ij} = 0$ according to the KKT condition. Therefore, we obtain the dual form of optimal transport:

$$\begin{aligned} &\max_{\boldsymbol{f},\boldsymbol{g}} \mathcal{J}_{\mathrm{OT}} = \langle \boldsymbol{f}, \boldsymbol{a}\rangle + \langle \boldsymbol{g}, \boldsymbol{b}\rangle \\ &s.t.\ f_i + g_j \leq C_{ij} \end{aligned} \tag{17}$$

Specifically, we can adopt the c-**transform** via $g_j = \inf_{k\in[M]} \left(C_{kj} - f_k\right)$. Meanwhile the optimal transport can be transformed into the following convex optimization problem:

$$\mathcal{J}_{\mathrm{OT}} = \arg\max_{\boldsymbol{f}} \left[\sum_{i=1}^{N} f_i a_i + \sum_{j=1}^{N} \left[\inf_{k\in[N]} \left(C_{kj} - f_k\right) \right] b_j \right] \tag{18}$$

We can adopt commonly-used optimization methods (e.g., L-BFGS) to obtain the optimal solution on $\boldsymbol{f}$. After we obtain the optimal result on $\boldsymbol{f}^*$, we can obtain $\boldsymbol{s}$ accordingly:

$$s_{ij} = C_{ij} - f_i^* - \inf_{k\in[N]} \left(C_{kj} - f_k^*\right) \tag{19}$$

Since we set $a_i = b_j = \frac{1}{N}$, the matching results in π_{ij} can be obtained as:

$$\pi_{ij} = \begin{cases} \dfrac{1}{N}, & s_{ij} = 0 \\ 0, & s_{ij} > 0 \end{cases} \tag{20}$$

B. Proof of Proposition 2

Proposition 2. *Given the stochastic differential equations $d\boldsymbol{z}_t = f(\boldsymbol{z}_t, t)dt + g(t)d\boldsymbol{w}_t$ with the drift and diffusion terms, the mean $\boldsymbol{\mu}(t)$ and covariance $\boldsymbol{\Sigma}(t)$ can be formulated as:*

$$\begin{cases} \frac{d\boldsymbol{\mu}(t)}{dt} = \mathbb{E}[f(\boldsymbol{z}, t)] \\ \frac{d\boldsymbol{\Sigma}(t)}{dt} = \mathbb{E}\left[f(\boldsymbol{z}, t)(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] \\ \qquad\quad + \mathbb{E}\left[(\boldsymbol{z}(t) - \boldsymbol{\mu}(t)) f^\top(\boldsymbol{z}, t)\right] + \mathbb{E}[g^2(t)] \end{cases} \tag{21}$$

Proof. To start with, it is noticeable that the mean value of the diffusion term $d\boldsymbol{w}_t$ is 0. Therefore, it is easy to verify that $\frac{d\boldsymbol{\mu}(t)}{dt} = \mathbb{E}[f(\boldsymbol{z}, t)]$. Meanwhile, the covariance term can be figure out as:

$$\begin{aligned} d\boldsymbol{\Sigma}(t) &= \mathbb{E}[d[(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top]] \\ &= \mathbb{E}[d(\boldsymbol{z} - \boldsymbol{\mu})(\boldsymbol{z} - \boldsymbol{\mu})^\top + (\boldsymbol{z} - \boldsymbol{\mu})d(\boldsymbol{z} - \boldsymbol{\mu})^\top + d(\boldsymbol{z} - \boldsymbol{\mu})d(\boldsymbol{z} - \boldsymbol{\mu})^\top] \end{aligned} \tag{22}$$

To simplify the first term, we should notice that:

$$\begin{aligned} &\mathbb{E}\left[(d\boldsymbol{z}(t) - d\boldsymbol{\mu}(t))(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] \\ &= \mathbb{E}\left[(d\boldsymbol{z}(t) - \mathbb{E}\left[f(\boldsymbol{z}, t)\right]dt)(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] \\ &= \mathbb{E}\left[d\boldsymbol{z}(t)(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] \end{aligned} \tag{23}$$

To simplify the second term, we also have the results as:

$$\begin{aligned} \mathbb{E}[d(\boldsymbol{z} - \boldsymbol{\mu})d(\boldsymbol{z} - \boldsymbol{\mu})^\top] &= \mathbb{E}\left[(g(t)d\boldsymbol{w}_t)(g(t)d\boldsymbol{w}_t)^\top\right] \\ &= \mathbb{E}[g^2(t)]dt \end{aligned} \tag{24}$$

Therefore, we have obtain the final solution:

$$\begin{aligned} d\boldsymbol{\Sigma}(t) &= \mathbb{E}\left[(d\boldsymbol{z}(t) - d\boldsymbol{\mu}(t))(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] \\ &\quad + \mathbb{E}\left[(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))(d\boldsymbol{z}(t) - d\boldsymbol{\mu}(t))^\top\right] + \mathbb{E}[g^2(t)]dt \\ &= \mathbb{E}\left[f(\boldsymbol{z}, t)(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))^\top\right] dt \\ &\quad + \mathbb{E}\left[(\boldsymbol{z}(t) - \boldsymbol{\mu}(t))\left(f(\boldsymbol{z}, t)\right)^\top\right] dt + \mathbb{E}[g^2(t)]dt \end{aligned} \tag{25}$$

□

C. Proof of Proposition 3

Proposition 3. *Given the Diverse Stochastic Differential Equations (DivSDE) as $d\boldsymbol{x}_t = \left(-\frac{1}{1-t}\right)\boldsymbol{x}_t dt + \eta\sqrt{\frac{2t}{1-t}}d\boldsymbol{w}_t$ with the initial data sample $\boldsymbol{z}_0$ and the noise level η, the probability of data distribution $\boldsymbol{z}_t$ is $p(\boldsymbol{x}_t) = \mathcal{N}((1-t)\boldsymbol{z}_t, \eta^2 t^2 \boldsymbol{I})$ at the time step t when $p(\boldsymbol{z}_0) = \mathcal{N}(\boldsymbol{z}_0, \boldsymbol{0})$.*

Proof. Adopting the Proposition 2, we can provide the equations on mean and covariance as below:

$$\begin{cases} \dfrac{d\boldsymbol{\mu}(t)}{dt} = \left(-\dfrac{1}{1-t}\right)\boldsymbol{\mu}(t) \\ \dfrac{d\boldsymbol{\Sigma}(t)}{dt} = \left(-\dfrac{2}{1-t}\right)\boldsymbol{\Sigma}(t) + \eta^2\dfrac{2t}{1-t} \end{cases} \tag{26}$$

The solutions are given as $\boldsymbol{\mu}(t) = (1-t)\boldsymbol{z}_i$ and $\boldsymbol{\Sigma}(t) = \eta^2 t^2 \boldsymbol{I}$. □

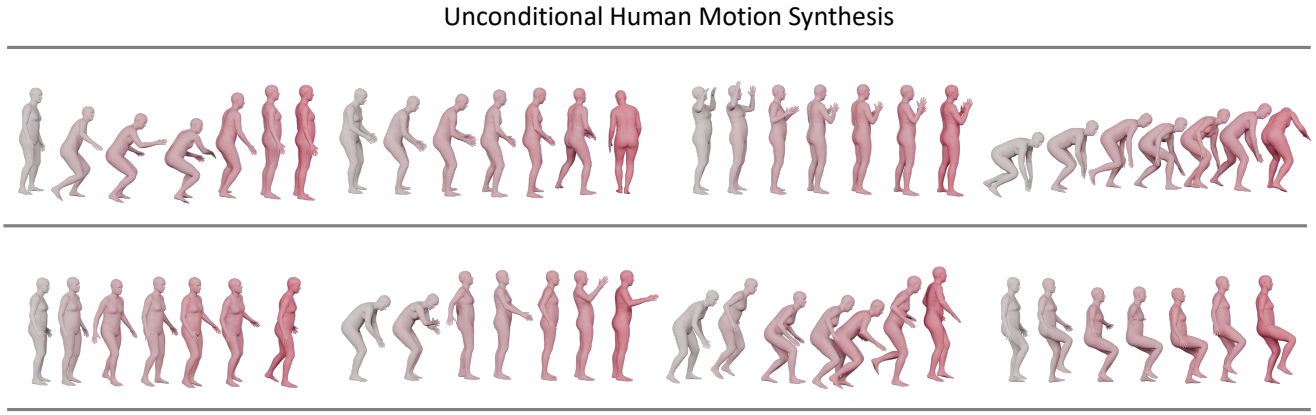

Figure 5. Qualitative results of DSDFM. We present more generated unconditional human motion sequences.

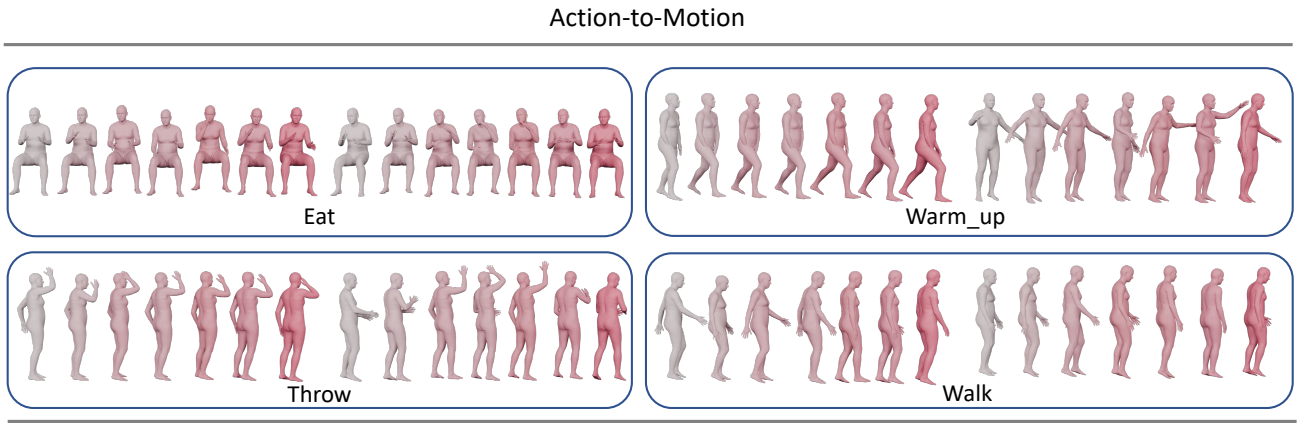

Figure 6. Qualitative results of DSDFM. We present the diverse human motion sequences under different actions.

D. Experiment Results

D.1. Metric Definitions

In this work, we use the following metrics to measure the performance of the proposed method for unconditional human motion synthesis and Action-to-Motion tasks.

Frechet Inception Distance (FID). FID calculates the distribution distance between the generated and real motions. FID is an important metric widely used to evaluate the overall quality of generated motions. The FID is calculated as:

$$\text{FID} = \|\mu_{gt} - \mu_{pred}\|^2 - \text{Tr}(\Sigma_{gt} + \Sigma_{pred} - 2(\Sigma_{gt}\Sigma_{pred})^{\frac{1}{2}}), \tag{27}$$

where Σ is the covariance matrix. Tr denotes the trace of a matrix. μ_{gt} and μ_{pred} are the mean of ground-truth motion features and generated motion features.

Kernel Inception Distance (KID). KID compares skewness as well as the values compared in FID [10], namely mean and variance. KID is known to work better for small and medium size datasets.

Precision, Recall. These measures are adopted from the discriminative domain to the generative domain [36]. Precision measures the probability that a randomly generated motion falls within the support of the distribution of real images, and is closely related with fidelity. Recall measures the probability that a real motion falls within the support of the distribution of generated images, and is closely related with diversity.

Accuracy. We use a pre-trained action recognition classifier [9] to classify human motions and calculate the overall recognition accuracy. The recognition accuracy indicates the correlation between the motion and its action type.

Diversity. Diversity measures the variance of the generated motions across all action categories. From a set of all generated motions from various action types, two subsets of the same size S_d are randomly sampled. Their respective sets of motion feature vectors $\{\mathbf{v}_1, \cdots, \mathbf{v}_{S_d}\}$ and $\left\{\mathbf{v}'_1, \cdots, \mathbf{v}_{S'_d}\right\}$ are extracted. The diversity of this set of motions is defined as:

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \| \mathrm{v}_i - \mathrm{v}'_i \|_2 \,. \tag{28}$$

where $S_d = 200$ is used in experiments.

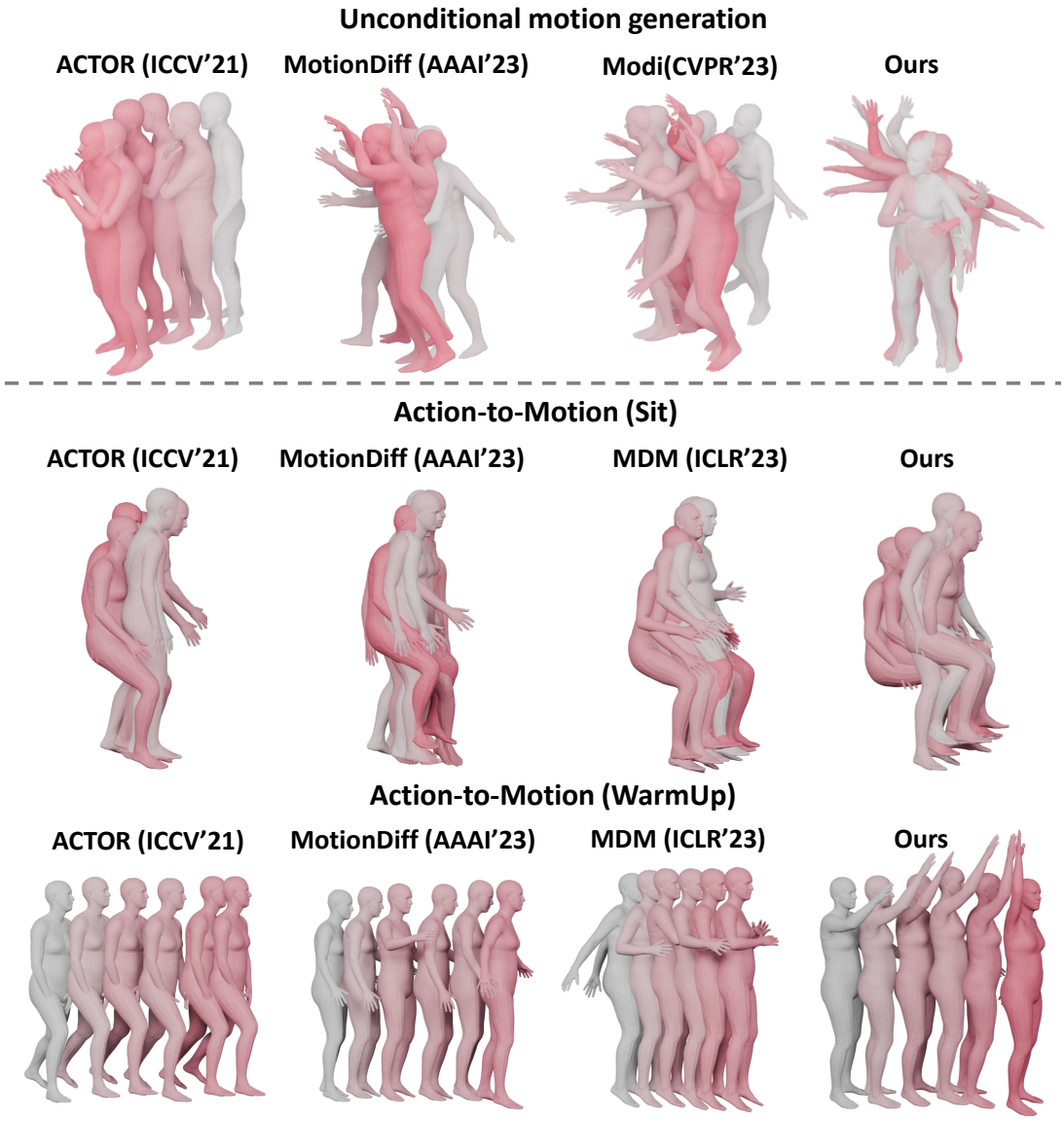

Figure 7. The qualitative comparison results of the state-of-the-art methods and our proposed DSDFM.

Multimodality. Different from diversity, multimodality measures how much the generated motions diversify within each action type. Given a set of motions with C action types. For c-th action, we randomly sample two subsets with the same size S_l, and then extract two subsets of feature vectors $\{\mathrm{v}_{c,1}, \cdots, \mathrm{v}_{c,S_l}\}$ and $\{\mathrm{v}'_{c,1}, \cdots, \mathrm{v}'_{c,S_l}\}$.The multimodality of this motion set is formalized as:

$$\text{Multimodality} = \frac{1}{C \times S_l} \sum_{c=1}^{C} \sum_{i=1}^{S_l} \left\| \mathrm{v}_{c,i} - \mathrm{v}'_{c,i} \right\|_2 . \quad (29)$$

where $S_l = 20$ is used in experiments

D.2. Additional Visualization Results

We provide additional visualization of human motion results in this section, which consists of the unconditional human motion synthesis and Action-to-Motion tasks.

Qualitative Analysis on Unconditional Human Motion Synthesis. Figure 5 visualizes a broader range of unconditional human motion sequences, effectively highlighting the remarkable diversity and high fidelity achieved by our proposed DSDFM. The visualization results demonstrate the remarkable ability of our method to produce diverse and realistic human motion sequences in unconditional human motion synthesis task.

Qualitative Analysis on Action-to-Motion. Figure 6 illustrates diverse human motion sequences across various action categories, providing evidence that our method is comparable under different action conditions.

Comparison with Other Methods. We provide more

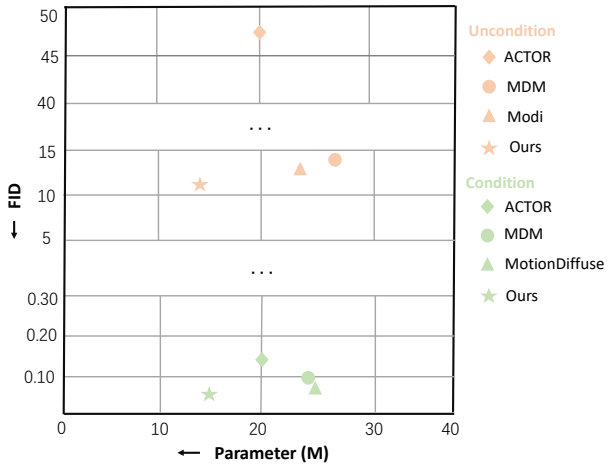

Figure 8. Comparison of the training parameter and the corresponding FID metric.

qualitative comparison of the state-of-the-art methods on human motion synthesis, i.e, unconditional motion generation and conditional motion generation under action labels (Action-to-Motion). As shown in Figure 7, we compare our method with the state-of-the-art methods. Under unconditional generation, the visual results of other methods show that the generated motion sequences tend to converge to static poses, resulting in a lack of diversity. Under action label conditional generation, some methods generate sequences that fail to meet the semantic requirements. The comparison results show that our method can achieve more diverse and accurate human motion sequences. More visualization results of our method can be seen in the supplementary video. These extensive results indicate that our method not only enables a significantly faster training process but also produces motion sequences with greater fidelity.

In addition, we visualize the comparison results of the training parameter and the corresponding FID metric. As shown in Figure 8. Our method achieves the best performance while utilizing the fewest training parameters. These results further underscore the effectiveness of the proposed approach.