TIMotion: Temporal and Interactive Framework for Efficient Human-Human Motion Generation Supplementary Material

Yabiao Wang[1,2*], Shuo Wang[2*], Jiangning Zhang[2], Ke Fan[3],
Jiafu Wu[2], Zhucun Xue[1], Yong Liu[1†]
[1]Zhejiang University [2]Youtu Lab, Tencent [3]Shanghai Jiao Tong University
https://aigc-explorer.github.io/TIMotion-page/

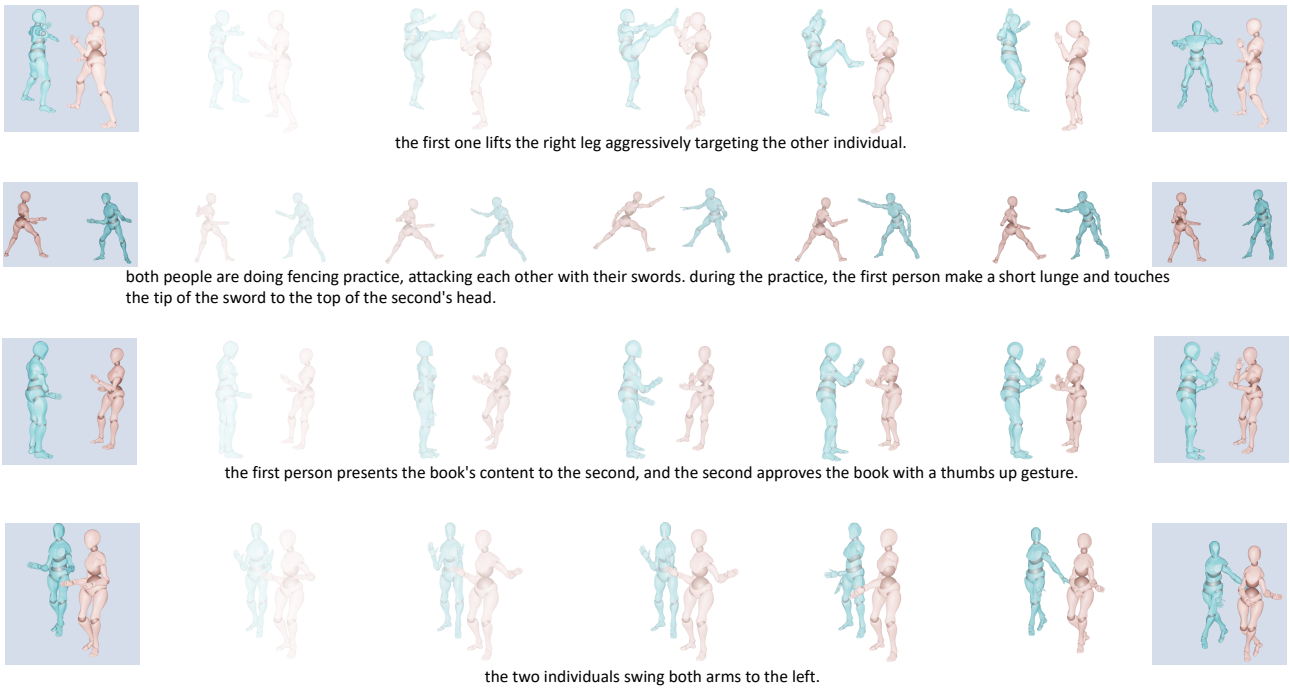

Figure 1. Qualitative results on the motion in-betweening task. The first and last frames are fixed. Darker colors indicate later frames.

Appendix

A. Theoretical Analyses

We perform gradient magnitude analysis on *separate modeling* (I) and our *causal interactive modeling* (II). Given that two single-person motion sequences X_a and X_b, the process of separate modeling is:

$$\begin{aligned} X_a^{out} &= \text{Softmax}(\frac{X_a W^Q (X_b W^K)^\top}{\sqrt{d}}) X_b W^V, \\ X_b^{out} &= \text{Softmax}(\frac{X_b W^Q (X_a W^K)^\top}{\sqrt{d}}) X_a W^V, \end{aligned} \tag{1}$$

where $\mathrm{W^Q}$, $\mathrm{W^K}$, and $\mathrm{W^V}$ are trainable weights. After causal interactive modeling, we can acquire X. Then we can obtain the final output as:

$$X_{out} = \text{Softmax}(\frac{(XW^Q)(XW^K)^\top}{\sqrt{d}}) XW^V. \tag{2}$$

For ease of analysis, we use the MSE Loss function:

$$\mathcal{L} = \frac{1}{2}\|X_a^{out} - Y_a\|_F^2 + \frac{1}{2}\|X_b^{out} - Y_b\|_F^2$$

The parameter gradient of separate modeling can be denoted as:

$$\nabla_W \mathcal{L}^{(I)} = \underbrace{(\mathbf{X}_a^{\text{out}} - \mathbf{Y}_a)}_{\Delta_a} \cdot \frac{\partial \text{ CrossAttn}}{\partial W_a} + \underbrace{(\mathbf{X}_b^{\text{out}} - \mathbf{Y}_b)}_{\Delta_b} \cdot \frac{\partial \text{ CrossAttn}}{\partial W_b}. \tag{3}$$

*Equal contributions.
†Corresponding author.

According to the properties of the matrix 2-norm: $\|A\|_F^2 = \text{Tr}(A^\top A)$, so we can get the F-norm of the parameter gradient as:

$$\|\nabla_W \mathcal{L}^{(I)}\|_F^2 = \|\Delta_a J_a + \Delta_b J_b\|_F^2. \tag{4}$$

Similarly, we can acquire the parameter gradient of our causal interactive modeling and its F-norm as:

$$\begin{aligned}\nabla_W \mathcal{L}^{(II)} &= [\Delta_a \mid \Delta_b] \cdot \frac{\partial \text{ SelfAttn}}{\partial W} \cdot \mathbf{X}_{out}^\top, \\ \|\nabla_W \mathcal{L}^{(II)}\|^2 &= \text{Tr}\left(\left([\Delta_a|\Delta_b] \cdot \frac{\partial \text{SelfAttn}}{\partial W} \cdot X_{out}^\top\right)^\top \left([\Delta_a|\Delta_b] \cdot \frac{\partial \text{SelfAttn}}{\partial W} \cdot X_{out}^\top\right)\right).\end{aligned} \tag{5}$$

where Tr represents the trace of a matrix.

As $\left([\Delta_a|\Delta_b] \cdot J_{self} \cdot X_{out}^\top\right)^\top = X_{out} \cdot J_{self}^\top \cdot [\Delta_a|\Delta_b]^\top$, we get the following equation:

$$\begin{aligned}&\text{Tr}\left(X_{out1} \cdot J_{self}^\top \cdot [\Delta_a|\Delta_b]^\top \cdot [\Delta_a|\Delta_b] \cdot J_{self} \cdot X_{out1}^\top\right) \\ &= \text{Tr}\left([\Delta_a|\Delta_b]^\top [\Delta_a|\Delta_b] \cdot J_{self} \cdot X_{out1}^\top X_{out1} \cdot J_{self}^\top\right),\end{aligned} \tag{6}$$

where J is the Jacobian matrix.

Assuming the input X_{out} is normalized and orthogonal ($X_{out}^\top X_{out} = I$), so we can obtain the final results:

$$\left\|\nabla_W \mathcal{L}^{(II)}\right\|_F^2 = \text{Tr}\left([\Delta_a \mid \Delta_b]^\top [\Delta_a \mid \Delta_b] \cdot J_{self}^\top J_{self}\right). \tag{7}$$

Then we do the following:

$$\begin{aligned}J_{\text{self}} &= \begin{bmatrix} J_a & J_{ab} \\ J_{ba} & J_b \end{bmatrix}, \\ J_{\text{self}}^\top J_{\text{self}} &= \begin{bmatrix} J_a^\top J_a + J_{ba}^\top J_{ba} & J_a^\top J_{ab} + J_{ba}^\top J_b \\ J_{ab}^\top J_a + J_b^\top J_{ba} & J_{ab}^\top J_{ab} + J_b^\top J_b \end{bmatrix},\end{aligned} \tag{8}$$

Next, we get the following two approximations:

$$J_{\text{self}}^\top J_{\text{self}} = \begin{bmatrix} J_a^\top J_a & J_a^\top J_{ab} \\ J_{ab}^\top J_a & J_b^\top J_b \end{bmatrix} + \sigma_{ab}(J_{ab}) + \sigma_{ba}(J_{ba}). \tag{9}$$

$$\|\nabla_W \mathcal{L}^{(II)}\|_F^2 > \text{Tr}\left(\Delta_a^\top \Delta_a J_a^\top J_a + \Delta_a^\top \Delta_b J_{ab}^\top J_a + \Delta_b^\top \Delta_a J_a^\top J_{ab} + \Delta_b^\top \Delta_b J_b^\top J_b\right) \tag{10}$$

where

$$\begin{aligned}\text{Tr}(\Delta_a^\top \Delta_a J_a^\top J_a) &= \|\Delta_a J_a\|_F^2, \\ \text{Tr}(\Delta_b^\top \Delta_b J_b^\top J_b) &= \|\Delta_b J_b\|_F^2, \\ \text{Tr}\left(\Delta_a^\top \Delta_b J_{ab}^\top J_a + \Delta_b^\top \Delta_a J_a^\top J_{ab}\right) &= 2 \cdot \text{Tr}\left(\Delta_a^\top \Delta_b J_{ab}^\top J_a\right).\end{aligned} \tag{11}$$

we apply the Cauchy-Schwarz and get:

$$\begin{aligned}\|\nabla_W \mathcal{L}^{(II)}\|_F^2 &> \text{Tr}\left(\Delta_a^\top \Delta_a J_a^\top J_a + \Delta_b^\top \Delta_b J_b^\top J_b + \Delta_a^\top \Delta_b J_{ab}^\top J_a + \Delta_b^\top \Delta_a J_a^\top J_{ab}\right) \\ &\geq \|\Delta_a J_a + \Delta_b J_b\|_F^2 + 2 \cdot \text{Tr}(\Delta_a^\top \Delta_b J_{ab}^\top J_a) \\ &> \|\Delta_a J_a + \Delta_b J_b\|_F^2 = \|\nabla_W \mathcal{L}^{(I)}\|_F^2\end{aligned} \tag{12}$$

Thus, our causal interactive modeling allows for faster model convergence.

B. More Experiments on InterX

To demonstrate the generalizability of our approach TIMotion, we perform corresponding experiments on another large-scale human-human motion generation dataset, InterX [6]. We maintain the same experimental setup as described in the paper. The results of comparative methods are directly borrowed from the InterX [6] paper except T2M* [1] and InterGen* [2]. The results of T2M* are taken from the open source repository of InterX [6] and the results of InterGen* are our own replication based on the unorganized training code provided by the authors of InterX and their open source validation code. Following established practices [2], each experiment is conducted 20 times, and the reported metric values represent the mean with a 95% statistical confidence interval. The results on InterX are shown in Tab. 1. In comparison to state-of-the-art (SOTA) approaches, our method, TIMotion, which incorporates various interaction mixing structures (including Transformer, Mamba, and RWKV), consistently outperforms others in terms of FID, R-Precision, Diversity, MM Dist, and MModality.

C. Algorithm of the motion in-betweening task

For the motion in-betweening task, we directly use the trained weights from the text-to-motion task. The overall inference process of motion in-betweening based on TIMotion is shown in Algorithm 1.

Algorithm 1 Inference of TIMotion on the Motion In-betweening Task

Input: Ground Truth of Motion Sequences for Two Individuals x^a and x^b, Length of the Sequence L, Ratio of Fixed Sequences α, Maximum Timestep of Diffusion T, Text Embedding c.
Output: Predicted Motion Sequences for Two Individuals $\hat{x}_0^a$ and $\hat{x}_0^b$.

$x_T^a \sim \mathcal{N}(0, I), \quad x_T^b \sim \mathcal{N}(0, I)$
for $t = T, \cdots, 1$ **do**
  $\hat{x}_0^a, \hat{x}_0^b$ = Diffusion(x_t^a, x_t^b, t, c)
  $\hat{x}_0^a[0 : L \cdot \alpha], \hat{x}_0^b[0 : L \cdot \alpha] = x^a[0 : L \cdot \alpha], x^b[0 : L \cdot \alpha]$
  $\hat{x}_0^a[L - L \cdot \alpha : L], \hat{x}_0^b[L - L \cdot \alpha : L] = x^a[L - L \cdot \alpha : L], x^b[L - L \cdot \alpha : L]$
  $\epsilon \sim \mathcal{N}(0, I)$
  $x_{t-1}^a = \sqrt{\overline{\alpha}_{t-1}}\hat{x}_0^a + \sqrt{1 - \overline{\alpha}_{t-1}}\epsilon$
  $x_{t-1}^b = \sqrt{\overline{\alpha}_{t-1}}\hat{x}_0^b + \sqrt{1 - \overline{\alpha}_{t-1}}\epsilon$
end for

Methods	R Precision↑			FID↓	MM Dist↓	Diversity→	MModality↑
	Top 1	Top 2	Top 3				
Real	$0.429^{\pm.004}$	$0.626^{\pm.003}$	$0.736^{\pm.003}$	$0.002^{\pm.0002}$	$3.536^{\pm.013}$	$9.734^{\pm.078}$	-
TEMOS [3]	$0.092^{\pm.003}$	$0.171^{\pm.003}$	$0.238^{\pm.002}$	$29.258^{\pm.0694}$	$6.867^{\pm.013}$	$4.738^{\pm.078}$	$0.672^{\pm.041}$
T2M [1]	$0.184^{\pm.010}$	$0.298^{\pm.006}$	$0.396^{\pm.005}$	$5.481^{\pm.3280}$	$9.576^{\pm.006}$	$5.771^{\pm.151}$	$2.761^{\pm.042}$
T2M* [1]	$0.325^{\pm.004}$	$0.487^{\pm.005}$	$0.593^{\pm.005}$	$3.342^{\pm.0572}$	$4.506^{\pm.020}$	$8.535^{\pm.055}$	$0.982^{\pm.054}$
MDM [5]	$0.203^{\pm.009}$	$0.329^{\pm.007}$	$0.426^{\pm.005}$	$23.701^{\pm.0569}$	$9.548^{\pm.014}$	$5.856^{\pm.077}$	$3.490^{\pm.061}$
MDM(GRU) [5]	$0.179^{\pm.006}$	$0.299^{\pm.005}$	$0.387^{\pm.007}$	$32.617^{\pm.1221}$	$9.557^{\pm.019}$	$7.003^{\pm.134}$	$3.430^{\pm.035}$
ComMDM [4]	$0.090^{\pm.002}$	$0.165^{\pm.004}$	$0.236^{\pm.004}$	$29.266^{\pm.0668}$	$6.870^{\pm.017}$	$4.734^{\pm.067}$	$0.771^{\pm.053}$
InterGen [2]	$0.207^{\pm.004}$	$0.335^{\pm.005}$	$0.429^{\pm.005}$	$5.207^{\pm.2160}$	$9.580^{\pm.011}$	$7.788^{\pm.208}$	$3.686^{\pm.052}$
InterGen* [2]	$0.400^{\pm.006}$	$0.585^{\pm.006}$	$0.695^{\pm.006}$	$0.475^{\pm.0305}$	$3.800^{\pm.020}$	$9.095^{\pm.055}$	$\underline{2.657}^{\pm.090}$
TIMotion+transformer(ours)	$\underline{0.412}^{\pm.004}$	$\underline{0.601}^{\pm.004}$	$\mathbf{0.714}^{\pm.003}$	$0.385^{\pm.0218}$	$\mathbf{3.706}^{\pm.015}$	$\mathbf{9.191}^{\pm.092}$	$2.437^{\pm.069}$
TIMotion+mamba(ours)	$\mathbf{0.414}^{\pm.005}$	$\mathbf{0.607}^{\pm.004}$	$\underline{0.713}^{\pm.003}$	$\underline{0.348}^{\pm.0170}$	$\mathbf{3.706}^{\pm.014}$	$9.095^{\pm.058}$	$\mathbf{2.779}^{\pm.083}$
TIMotion+RWKV(ours)	$0.411^{\pm.005}$	$0.597^{\pm.006}$	$0.707^{\pm.004}$	$\mathbf{0.261}^{\pm.0140}$	$\underline{3.737}^{\pm.015}$	$\underline{9.112}^{\pm.079}$	$2.475^{\pm.075}$

Table 1. **Quantitative evaluation on the InterX [6] test set.** We run the evaluations 20 times. ± indicates a 95% confidence interval. **Bold** indicates the best result, while underline refers to the second best. The results of comparative methods are directly borrowed from the InterX [6] paper except T2M* [1] and InterGen* [2]. The results of T2M* are taken from the open source repository of InterX [6] and the results of InterGen* are our own replication based on the unorganized training code provided by the authors of InterX and their open source validation code.

D. More Qualitative Results

Human-Human motion generation. We provide the supplemental demo named **demo.mp4**.

Motion Editing. We provide qualitative results on the motion in-betweening task in Fig. 1. Our method achieves smooth and natural transitions between the conditioned and generated motions while complying with the text.

E. Metrics Computation

Frechet Inception Distance (FID): Features are extracted from generated motions and real motions. Subsequently, FID is calculated by comparing the feature distribution of the generated motions with that of the real motions. FID serves as a crucial metric extensively utilized to assess the overall quality of the synthesized motions.

R Precision: For each generated motion, a description pool is created consisting of its ground-truth text description and 31 randomly chosen mismatched descriptions from the test set. Next, the Euclidean distances between the motion and text features of each description in the pool are computed and ranked. We then calculate the average accuracy at the top-1, top-2, and top-3 positions. If the ground truth entry appears among the top-k candidates, it is considered a successful retrieval; otherwise, it is deemed a failure.

MM Dist: MM distance is calculated as the mean Euclidean distance between the motion feature of each generated motion and the text feature of its corresponding description in the test set.

Diversity: Diversity measures the variance of the generated motions. From the entire set of generated motions, two subsets of the same size S_d are randomly sampled. Their respective sets of motion feature vectors $\{\mathbf{v}_1, ..., \mathbf{v}_{S_d}\}$ and $\{\mathbf{v}'_1, ..., \mathbf{v}_{S'_d}\}$ are extracted. The diversity of this set of motions is defined as

$$\text{Diversity} = \frac{1}{S_d} \sum_{i=1}^{S_d} \parallel \mathbf{v}_i - \mathbf{v}'_i \parallel_2 . \tag{13}$$

$S_d = 300$ is used in experiments.

MModality: MModality measures how much the generated motions diversify within the same text. Given a set of motions corresponding to a specific text, two subsets of the same size S_l are randomly sampled. Their respective sets of motion feature vectors $\{\mathbf{v}_{c,1}, ... , \mathbf{v}_{c,S_l}\}$ and $\{\mathbf{v}'_{c,1}, ..., \mathbf{v}'_{c,S_l}\}$ are extracted. The MModality of this motion set is formalized as

$$\text{Multimodality} = \frac{1}{C \times S_l} \sum_{c=1}^{C} \sum_{i=1}^{S_l} \left\| \mathbf{v}_{c,i} - \mathbf{v}'_{c,i} \right\|_2 . \tag{14}$$

$S_l = 100$ is used in experiments.

F. Limitation

Limited by the variety of two-person datasets, we demonstrated the effectiveness of TIMotion mainly on the text-to-motion task. In the future, we will validate our method on more tasks based on newly released datasets. Moreover, our proposed TIMotion effectively models the motion relationship between the two individuals, but the modeling of motion relationships between three or more people has not been explored yet. Related researchers in the community are encouraged to explore more on motion modeling among people and TIMoiton may provide some new insight for the community.

References

[1] 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*, pages 5152–5161, 2022. 2, 3

[2] Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, pages 1–21, 2024. 2, 3

[3] Mathis Petrovich, Michael J Black, and Gül Varol. Temos: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision*, pages 480–497. Springer, 2022. 3

[4] Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. Human motion diffusion as a generative prior. *arXiv preprint arXiv:2303.01418*, 2023. 3

[5] G Tevet, S Raab, B Gordon, Y Shafir, D Cohen-Or, and AH Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022. 3

[6] Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22260–22271, 2024. 2, 3