FreeMotion: A Unified Framework for Number-free Text-to-Motion Synthesis (Supplementary Material)

Ke Fan[1], Junshu Tang[1], Weijian Cao[2], Ran Yi[1⋆], Moran Li[2], Jingyu Gong[1], Jiangning Zhang[2], Yabiao Wang[3,2], Chengjie Wang[1,2], and Lizhuang Ma[1,4,5⋆]

[1] Shanghai Jiao Tong University
{slipperyfrank, tangjs, ranyi, gongjingyu, lzma}@sjtu.edu.cn
[2] Tencent Youtu Lab
weijiancao@tencent.com, moranli.aca@gmail.com, {caseywang, jasoncjwang}@tencent.com
[3] Zhejiang University
[4] East China Normal University
[5] MoE Key Lab of Artificial Intelligence, Shanghai Jiao Tong University
https://VankouF.github.io/FreeMotion

In the following, we provide additional implementation details, where we introduce the metrics, settings, and the usage of LLM for the separation of motion description, and the details of training as well as the manner of visualization. Then, we show the additional experiments, which include more qualitative results, comparison between self-attention and cross-attention mechanism, and an user study. We also provide a demo video for visualizations of the generated motions.

1 Additional Implementation Details

1.1 Metrics

We utilize the same evaluation metrics as InterGen. (1) FID: measuring the latent distribution distance between the generated dataset and the real dataset. (2) R Precision: measuring the text motion matching, indicates the probability that the real text appears in the Top-k (1, 2, and 3) after sorting. (3) Diversity: measuring latent variance. (4) Multimodality (MModality): measuring diversity within the same text. (5) Multi-modal distance (MM Dist): measuring the distance between motions and text features.

1.2 Settings

We employ a frozen *CLIP-ViT-L-14* model as the text encoder. The number of diffusion timesteps is set to 1,000 during training and the DDIM [6] sampling strategy with 50 timesteps is applied in the inference stage. For **number-free**

⋆ Corresponding Authors.

motion generation, the batch size is set to 80 and 30 for the first and the second training stage on each GPU. The epoch numbers are 2,500 and 1,000 separately for the two stages. Both training stages are trained with 1e-4 learning rate and 2e-5 weight decay. For **spatial control**, the epoch number and learning rate are set to 1,000 and 1e-5. Other hyper-parameters remain the same as aforementioned. All experiments are trained on eight Tesla V100 GPUs.

1.3 LLM for Single-person Motion Description.

To facilitate the generation module in synthesizing single-person motion, it is essential to provide a single-person motion description as input during the training phase. To accomplish this, given a multi-person interactive motion description, we leverage the capabilities of the Large Language Model (LLM) [5] and carefully design the prompt to generate N motion descriptions for each single individual. This process involves two steps: (1) **Separation**, wherein the LLM divides the interaction description into single-person motion descriptions, and (2) **Assignment**, wherein the split texts are assigned to their respective individuals.

Separation. LLM is a conversational model built on a large language model that can facilitate natural language conversations and generate corresponding responses. The responses generated by LLM are often directly influenced by the information and expressions provided in the prompt. For our task, inspired by [7], we meticulously designed effective prompts through empirical validation. The prompt and some typical results are shown in Tab. 2.

Table 1: We use CA to represent the cross-attention mechanism and the SA to represent the self-attention mechanism.

Method	FID ↓
CA	7.03
SA	**6.74**

Assignment. To assign each single-person description to the corresponding motion, manual allocation is not feasible due to the high cost involved. Therefore, we employ a feature extraction model developed by [2], which is trained on a large-scale single-person motion dataset with a text annotation (HumanML3D [2]) by contrastive learning, to aid us in the assignment process. This model aims to bring paired text-motion features closer during training while pushing unpaired ones farther away. We extract the features of each text and motion using this model and calculate the Euclidean distance between them. As we only have a large-scale text-motion dataset for two individuals, given two text descriptions t_1, t_2, and two motions m_1, m_2, we calculate a 2×2 distance matrix as shown in Fig. 1(a). However, directly comparing the feature space distance between (t_1, m_1), (t_1, m_2) and (t_2, m_1), (t_2, m_2) to assign motion to text may

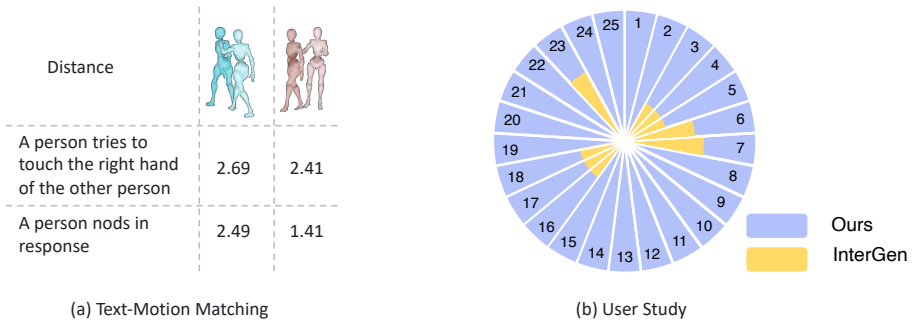

Fig. 1: The left (a) demonstrates a case of the calculated Euclidean distance between text features and motion features. The right (b) shows the results of our user study, and each sector represents the corresponding question. The larger the area, the more people choose the corresponding model.

result in one-to-many mapping, *e.g.,* when $d_{t_1 m_1} < d_{t_1 m_2}$ and $d_{t_2 m_1} < d_{t_2 m_2}$, both t_1 and t_2 are assigned to m_1. To avoid the one-to-many assignment, we have designed a simple rule shown in Algorithm. 1, to ensure the mapping from text to motion is a one-to-one mapping.

1.4 Mixed Texts for Training

As analyzed in the experimental section of our main paper, incorporating both our split single-person text descriptions and high-quality interaction descriptions during the training phase can substantially enhance the model's performance. To achieve this, we input the corresponding single-person text and global interaction text for each individual into the frozen CLIP model to extract their respective text features. We then add these two features and use the resultant feature as a condition to guide the model during subsequent training to generate the desired motion for that individual.

1.5 Visualizations

Once the model predicts the joints, we obtain the corresponding SMPL [4] parameters through SMPL model fitting. To enhance the visual appeal of the display, we download the character from Mixamo [1] and retarget it through the blender plug-in. Please refer to our accompanied demo video for visualizations.

2 Additional Experiments

2.1 Additional Qualitative Results.

In this section, we provide additional comparisons of our generated single and two-person motions with our revised InterGen [3]. The results shown in Fig. 2

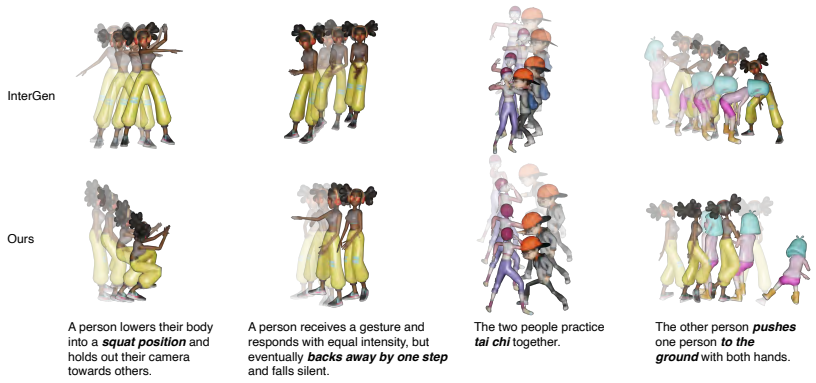

Fig. 2: Comparison with InterGen* on single and two-person motion generation. For single-person motion, we generate it with our re-annotated single description. For two-person motion, we further leverage the original interactive descriptions. For better visualization, some pose frames are shifted to prevent complete overlap.

demonstrate that our method outperforms InterGen in both single and multi-person motion generation.

2.2 Comparison between Self-Attention and Cross-Attention

Compared with cross-attention, self-attention extracts not only the interactive relations between noise motion and motion conditions but also the interaction between different motion conditions. As shown in Tab 1, where SA achieves a better FID, which means that compared with cross-attention, self-attention is more flexible for modeling the interaction between motions regardless of the number of motions in the condition.

2.3 User Study

We conduct a user study to obtain the user's subject evaluation between ours and InterGen for the generation of both single and two-person motions. We generate 25 results using our method, and 25 results using InterGen under the same set of input texts, and present these results to more than 50 participants. We ask them to select their preference according to the comprehensive consideration of text-motion consistency and fidelity. As shown in Fig. 1(b), our method achieves an overwhelming majority of votes in all cases, proving that our carefully designed model architecture can better adapt to motion generation for any number of people.

Algorithm 1 Text Motion Matching

Require: A list $L = \{(d^i_{t_1m_1}, d^i_{t_1m_2}, d^i_{t_2m_1}, d^i_{t_2m_2})\}, i \in \{1, ..., N\}$, where d represents the distance of corresponding text feature and motion feature, N represents the total interactive descriptions.

for $d_{t_1m_1}, d_{t_1m_2}, d_{t_2m_1}, d_{t_2m_2}$ in L **do**
  $p_{t_1m_1}, p_{t_1m_2} = \frac{d_{t_1m_2}}{d_{t_1m_1}+d_{t_1m_2}}, \frac{d_{t_1m_1}}{d_{t_1m_1}+d_{t_1m_2}}$
  $p_{t_2m_1}, p_{t_2m_2} = \frac{d_{t_2m_2}}{d_{t_2m_1}+d_{t_2m_2}}, \frac{d_{t_2m_1}}{d_{t_2m_1}+d_{t_2m_2}}$
  if $p_{t_1m_1} > p_{t_1m_2}$ **then**
    if $p_{t_2m_1} \leq p_{t_1m_2}$ **then**
      Match(t_1,m_1), Match(t_2,m_2)
    else
      if $p_{t_1m_1} \geq p_{t_2m_1}$ **then**
        Match(t_1,m_1), Match(t_2,m_2)
      else
        Match(t_1,m_2), Match(t_2,m_1)
      end if
    end if
  else
    if $p_{t_2m_2} \leq p_{t_2m_1}$ **then**
      Match(t_1,m_2), Match(t_2,m_1)
    else
      if $p_{t_1m_1} \geq p_{t_2m_1}$ **then**
        Match(t_1,m_1), Match(t_2,m_2)
      else
        Match(t_1,m_2), Match(t_2,m_1)
      end if
    end if
  end if
end for
return

Table 2: Example of the designed prompt and a few cases for separating the interactive description into the single-person motion texts by LLM.

Separation Process (two people as an example)
Instruction: two people greet each other with a handshake while holding their cards in the left hand. Given the instruction, generate the descriptions according to the following rules and examples. Each task plan should completely reflect an entire process of actions described in the instruction. [**start of rules**] 1. Be plausible. Do not generate uncommon interactions. 2. Use one sentence to describe what action should person 1 do and one sentence to describe what action should person 2 do according to the text instruction at the beginning of the task plan. **IMPORTANT**: 1. the sentence starts from 'text 1:' describing the action of person 1 from the perspective of person 1 and the sentence starts from 'text 2:' describing the action of person 2 from the perspective of person 2. Sentences should NOT contain words like 'person 1' or 'person 2', use 'a person' to refer to himself in the sentence and 'others' to refer to others. 3. Please note that if a description only contains the actions of one person, it means that the second person stands still. In this case, please give the description of the second person the default output as "A person stands still. [**end of rules**] [**start of an example 1**] Instruction: two people greet each other with a handshake, while holding their cards in the left hand. Output: [Text 1] a person make a handshake with others using his right wrist, while holding his cards in the left wrist. [Text 2] a person make a handshake with others using his right wrist, while holding his cards in the left wrist. [**end of an example 1**] [**start of an example 2**] Instruction: one person moves ahead, lowering the arm. Ouput: [Text 1] A person drops their arm and moves a step forward. [Text 2] A person stands still. [**end of an example 2**]
User: the first one lowers the arm and moves ahead. **LLM**: Output [Text 1] A person lowers their arm and takes a step forward. [Text 2] A person stands still. **User**: one person nods when the other person tries to touch the right hand. **LLM**: Output: [Text 1] A person nods in response as the other person reaches out to touch their right hand. [Text 2] A person tries to touch the right hand of the other person while observing their reaction.

References

1. Adobe: Https://www.mixamo.com/ 3
2. Guo, C., Zou, S., Zuo, X., Wang, S., Ji, W., Li, X., Cheng, L.: Generating diverse and natural 3d human motions from text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5152–5161 (2022) 2
3. Liang, H., Zhang, W., Li, W., Yu, J., Xu, L.: Intergen: Diffusion-based multi-human motion generation under complex interactions. arXiv preprint arXiv:2304.05684 (2023) 3
4. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023) 3
5. OpenAI: Https://openai.com/blog/chatgpt 2
6. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020) 1
7. Wang, Z., Wang, J., Lin, D., Dai, B.: Intercontrol: Generate human motion interactions by controlling every joint. arXiv preprint arXiv:2311.15864 (2023) 2