FreeDance: Towards Harmonic Free-Number Group Dance Generation via a Unified Framework

A. Overview

The supplementary material includes the subsequent components.

- Details of Methodology and Experiment.
  - Insights for motion token residual.
  - Auxiliary losses.
  - Explanation on the evaluation metrics.
- Additional Visualization Results
  - Visualization of the ablation studies.
  - Results of extending the maximum person number to 5.
- Demo Page and Demo Video

B. Details of Methodology and Experiment

B.1. Insights for motion token residual

We introduce the motion token residual in Section 3.3 as the key component to enhance the cross-modality alignment.

As defined in our Cross-modality Residual Alignment Module (CRAM), motion residual is the successive subtraction of motion tokens along the temporal dimension. Unlike in raw motion space, where consecutive frames sometimes lack completeness and easily become jittering, motion tokens provide a temporally more compact alternation. The residual of motion token is the difference of body parts dynamics, which is fundamentally the direction change in token map feature space. Tensors with smaller distances are more likely to have closer semantic meaning, which, in our case, is similar motion. Thus, the motion token residual indicates the amplitude of movements, which is considered more tightly correlated with music features.

B.2. Auxiliary losses

In Section 3.4, we mention the additional losses we used for stage 2: masked token modeling training. It contains the velocity loss $\mathcal{L}_{\text{vel}}$, the forward kinetics reconstruction loss $\mathcal{L}_{\text{joint}}$, and the foot contact loss $\mathcal{L}_{\text{contact}}$. Their definition follows previous works [3, 8].

$$\mathcal{L}_{\text{joint}} = \frac{1}{N}\sum_{n=1}^{N} \| \operatorname{FK}(\hat{d}^n) - \operatorname{FK}(d^n)\|_2^2. \quad (1)$$

$$\mathcal{L}_{\text{vel}} = \frac{1}{N-1}\sum_{n=1}^{N} \|(d^{n+1} - d^n) - (\hat{d}^{n+1} - \hat{d}^n)\|_2^2. \quad (2)$$

$$\mathcal{L}_{\text{contact}} = \frac{1}{N-1}\sum_{n=1}^{N} \|(\operatorname{FK}(\hat{d}^{n+1}) - \operatorname{FK}(\hat{d}^n)) \cdot \hat{\mathbf{b}}^n\|_2^2. \quad (3)$$

Here, N represents the total frames in a motion sequence. n denotes the temporal index of the frame. FK is the forward kinematics operator. $\hat{\mathbf{b}}^n$ is the binary foot contact label. We utilize these auxiliary losses to further improve the smoothness and naturalness of the decoded motions.

B.3. Explanation on the evaluation metrics

In Section 4.1, we briefly introduce the evaluation metrics used for our quantitative analysis. After a board survey, we found the same metrics (i.e., FID, Diversity) used in solo dance, group dance, and motion generation communities are not exactly unanimous. The main difference lies in the motion feature extraction process. We introduce the reason for our metric selection here, with the fundamental purpose of faithfully modulating the data distribution.

- **FID:** Fréchet Inception Distance (FID) is a similarity measurement for two distributions. Although quite a portion of dance generation works use kinetics and geometry features to define the distribution and report FID_k, FID_g results, we argue that these manually defined features have relatively large information loss through the feature extracting process. As mentioned in EDGE [8], we also observe fluctuation of FID_k and FID_g in the training process and a relatively low correlation between the score and the visualization results. Therefore, we follow Choreomaster [1], Dancing with Music [2] in dance generation, and Mmm [6] in the motion generation domain, using the mixed AIST++ and AIOZ-GDANCE dataset to train a motion auto-encoder. Then use this self-supervised learning-based feature to calculate the distance between ground truth and generated motion distributions.
- **BAS:** Beat Align Score measures the temporal alignment of music and dance. We follow previous implementations [3, 4, 7], using Librosa [5] to extract the music beat

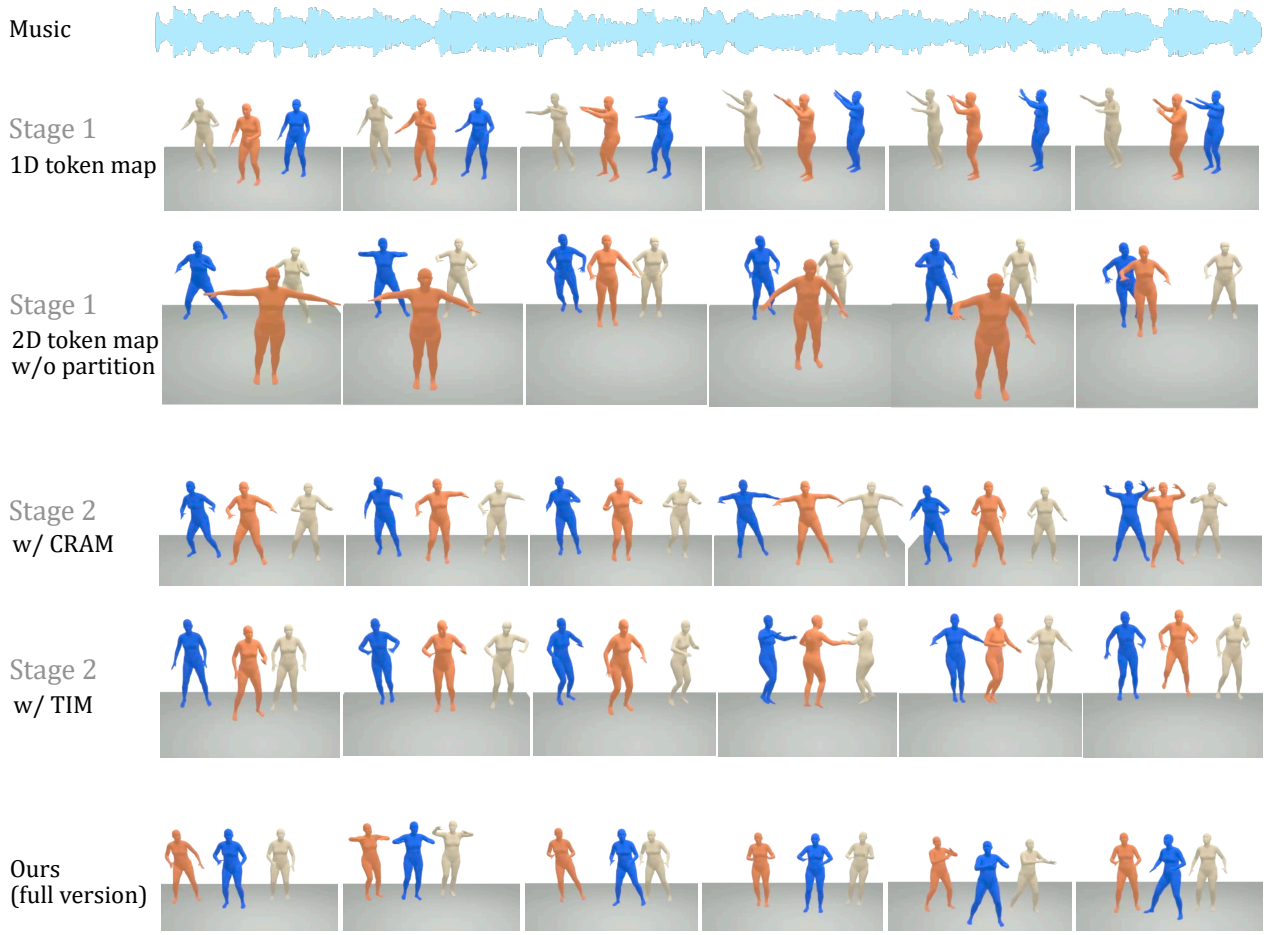

Figure 1. Given the same music segments, the results generated by our full version framework versus ablation versions.

features and calculate the motion beat according to the difference in joint positions between adjacent frames. A higher score indicates better cross-modality synchronization.

- **Diversity:** Diversity score is the distance between generated motion features conditioned on different music inputs. Similar to the FID calculation, we follow [1, 2, 6] to use the same motion autoencoder as the feature extractor. The more variation the generated dance has, the higher the diversity score is.
- **Wins:** The Wins score reflects subjective user preferences in our study. We collect two sets of samples: one generated by our FreeDance framework and the other consisting of randomly selected samples from comparison methods. The results indicate the percentage of our samples preferred by users.

C. Additional Visualization Results

C.1. Visualization of the ablation studies

We demonstrate the dance keyframes of our ablation studies. As stated in Section 4.2, we first show the effectiveness of our 2D token map with the person-number-dependent quantization strategy. The results in Figure 1 show the 1D token map leads to slow and freeze motions. 2D token map w/o person-number-dependent partition cannot avoid the leakage of static pose to generated sample. We then show the benefits of adding the Cross-modality Residual Alignment Module (CRAM) and Temporal Interaction Module (TIM). CRAM improves the coherence between dance and music conditions, preserving group harmony, while TIM increases individual motion variation and group coordination.

C.2. Results of extending the maximum person number to 5

Although our main experiment is conducted by setting the maximum person number to three due to the unequal person-number distribution of the dataset, our framework can generalized to user-specified free-number dancers. We also train our framework using one-person data from AIST++ and two-to-five-person data from AIOZ-GDANCE then show the dance key frames of four and five dancers in Figure 2. Please further refer to the demo video for dynamic visualization.

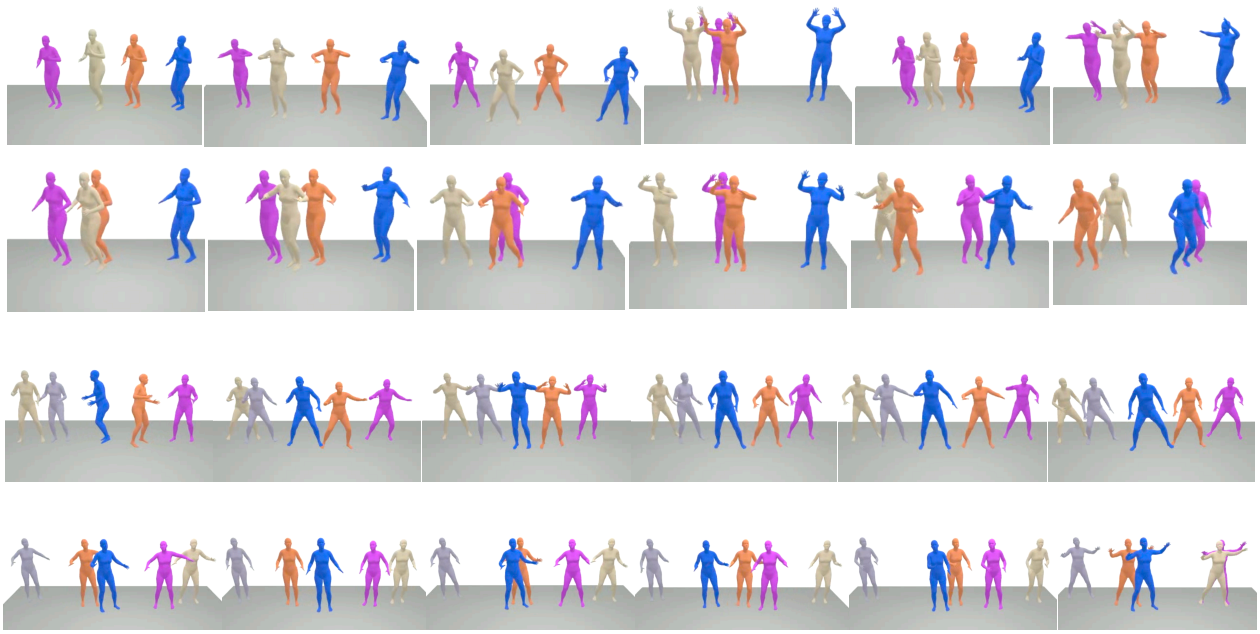

Figure 2. The results of four and five dancers generated by our framework.

D. Demo Page and Demo Video

We provide a demo page (see the attached index.html file in the demo_page folder), that presents a concise recap of our framework. The page also includes an embedded demo video, showing our framework design (0'00–0'44). The video results of the comparison between our method and other state-of-the-art approaches (0'44–1'33), the ablation results (1'33–1'40), and the additional free-number results generated by our framework (1'40–2'17).

References

[1] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 1, 2

[2] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 1, 2

[3] Ronghui Li, YuXiang Zhang, Yachao Zhang, Hongwen Zhang, Jie Guo, Yan Zhang, Yebin Liu, and Xiu Li. Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1524–1534, 2024. 1

[4] Zhenye Luo, Min Ren, Xuecai Hu, Yongzhen Huang, and Li Yao. Popdg: Popular 3d dance generation with popdanceset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26984–26993, 2024. 1

[5] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. *SciPy*, 2015:18–24, 2015. 1

[6] Ekkasit Pinyoanuntapong, Pu Wang, Minwoo Lee, and Chen Chen. Mmm: Generative masked motion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1546–1555, 2024. 1, 2

[7] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3d dance gpt with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[8] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 1