TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration (Supplementary Material)

Kehong Gong[1*] Dongze Lian[1*] Heng Chang[2] Chuan Guo[2]
Zihang Jiang[1] Xinxin Zuo[2] Michael Bi Mi[2] Xinchao Wang[1†]
[1] National University of Singapore [2] Huawei Technologies Co., Ltd.

Abstract

This supplementary material provides more details on the following aspects of our study: i) The dataset we used; ii) The evaluation metrics we employed; iii) The impact of mixed data for shared motion token space; iv) The collected data distribution; v) The effect of music-text fusion weight; vi) The reason why our dance has less freeze issue; vii) More visualizations of our results.

1. Detail of Dataset

For the music2dance dataset, we employ the AIST++ dataset [5], which contains 30 subjects and 10 dance genres. There are 992 pieces of 3D human pose sequence, of which 952 are used for training and the rest are used for evaluation.

For the text2motion dataset, we employ the HumanML3D [2] dataset, which is a large-scale 3D human motion dataset that covers a broad range of human actions such as locomotion, sports, and dancing. It consists of 14,616 motions and 44,970 text descriptions. Each motion clip comes with at least 3 descriptions. For the joint training of both datasets, we sample the motions with 60 frames per second (FPS) to keep the time consistency with the AIST++ dataset, resulting in duration ranges from 2 to 10 seconds.

To evaluate the generalization ability of our method, we also collected a new dataset of music clips from YouTube that are not included in AIST++. This dataset consists of 82 clips with a total duration of 53 minutes, which is eight times larger than the AIST++ test set. The clips cover various styles and content of music, which are out of the distribution of AIST++. In detail, our data are popular music collected from YouTube, which covers a variety of styles such as Glitch hop, Electro house, rock, future bass, indie pop, and R&B. By contrast, AIST++ uses pure dance music from Old School (Break, Pop, Lock, and Waack) and New School (Middle Hip-hop, LA-style Hip-hop, House, Krump, Street Jazz, and Ballet Jazz) genres. Additionally, we selected in-the-wild music based on popularity, such as "Faded," "Beat It," "Coincidence," "Baby," "Poker Face," "Despacito," "Panama," "Love Story," and others, with millions of plays. Furthermore, we provide a t-SNE feature distribution diagram (Figure 4) to demonstrate the diversity and distinctiveness of our dataset compared to AIST++.

2. Evaluation Metrics

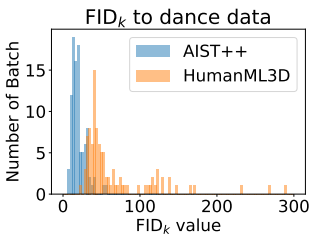
We follow FACT [5] and Bailando [9] to quantitatively measure the quality of generated dances, the diversity of motions and the beat alignment of the music and the generated motions. In concrete, for the dance quality, we calculate the Fréchet Inception Distances (FID) [4] between the generated 3D dance and all motions of the AIST++ dataset on kinetic features [7] (denoted as 'k') and geometric features [6] (denoted as 'g') extracted by [1] to measure the quality of generated dances. We also follow [5] to calculate the average feature distance of generated motion to measure the diversity of motions. The average distance between the music beat and its closest dance beat is defined as the Beat Align Score as follows

$$\frac{1}{|B^m|}\sum_{b^m\in B^m}\exp\left\{-\frac{\min_{b^d\in B^d}\|b^d-b^m\|^2}{2\sigma^2}\right\}, \quad (1)$$

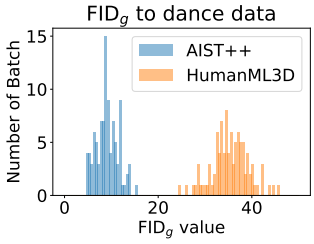
where B^d and B^m are the dance beats and music beats, respectively. σ is a normalized parameter that we set to be $\sigma = 3$ in our experiments.

For the text2motion quality, we follow the same setting suggested by TM2T [3]: R-precision and Multimodal-Dist quantify the relevancy between the generated motions and the input prompts; FID computes the distance between the generated and ground truth distributions (in latent space); Diversity evaluates the variation of the generated motions; and MultiModality estimates the variance for a single prompt

*Equal contribution: gongkehong@u.nus.edu, dongze@nus.edu.sg
†Corresponding author: xinchao@nus.edu.sg



(a) Difference in FID$_k$.



(b) Difference in FID$_g$.

Figure 1: FID$_k$ and FID$_g$ with difference batches in Experiment A.

We also introduce two new evaluation metrics: Percentage of Freezing Frame (PFF) and Motion Prediction Distance (MPD). PFF measures the degree of freezing in the generated dance, while MPD assesses the coherence of frames when text is integrated.

3. The Impact of Mixed Data

As mentioned in the main text, a direct combination of the music2dance (AIST++ [5]) and text2motion (HumanML3D [2]) in the motion space might be sub-optimal for training because the motions from these two datasets fall in completely different spaces. In contrast, we project the motions into a consistent and shared latent space with a human motion VQ-VAE architecture. To show the effectiveness of the proposed method quantitatively, we design two experiments as follows.

- Experiment A: we random sample 100 batches of data (same size as AIST++ test set) from both datasets, and measure the FID between the random batch and the whole dance data.
- Experiment B: we sample 30% of the original data from both datasets and train them with a human motion VQ-VAE of different downsample rates (4, 8, 16, 32).

In experiment A, Figure 1 shows the distribution of FID results from both datasets. From Figure 1, we can observe that there is a distinct difference between the two datasets on geometric feature, and a small overlap in kinetic feature.

In experiment B, from the Figure 2, we have the following three findings: i) Figure 2 (a) and Figure 2 (b) show that the tokens used of each dataset will be increasing with the training epoch. ii) In Figure 2 (c), the shared token number is also increasing together with it from both dataset. iii) The lower the downsample rate, the higher the used token number and shared token number, with smaller reconstruction loss (val loss). Consider that the lower the downsample rate, the longer the tokenized sequence for transformer in the second step of our pipeline. We choose downsample rate of 8, (a relatively small val loss, rich shared token number, and relatively short tokenized sequence length).

From Figure 3, we can see that both datasets almost share one codebook when motions are encoded with a VQ-VAE. Specifically, the total number of vectors contained in the codebook is 1024, 855 vectors and 912 vectors of which are used to construct the motions in AIST++ and HumanML3D, respectively. 846 vectors (98.9% in AIST++ and 92.8% in HumanML3D) are shared to generate the motion tokens, which is much better than the feature distance from Figure 1.

4. The Analysis of the Collected Dataset

To verify the domain gap between source music and wild music, We sample the music features extracted by the Librosa (used in framework training) and plot a t-SNE in Figure 4. Two music datasets lay on two different distributions with a few overlaps, which shows the generalization ability of our method. The inferior results in Table 1 (main text) compared with our mix training show that mix gains better generalization performance.

5. The Fusion Weight and Text2motion Results

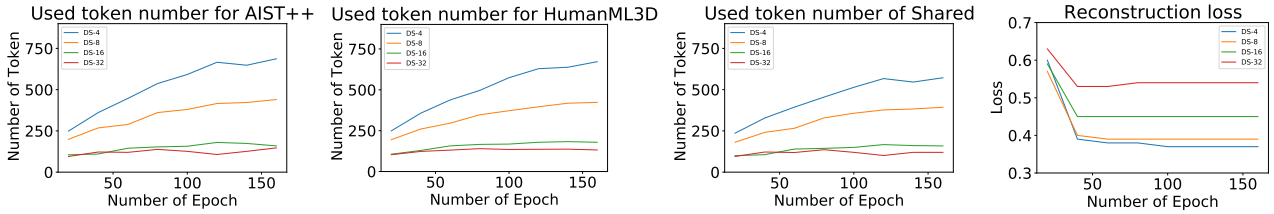
We further explore the effect of late fusion rate (LFR), as shown in Fig 5, with the increasing of LFR, the MM distance and Top 1 precision get worse. To balance the feature content, we choose late fusion rate of 0.8.

train	AIST++		mix data	
test	AIST++	wild	AIST++	wild
1	4.08 / 4.08	3.76 / 3.40	5.67 / 4.88	1.21 / 0.87
10	1.31 / 1.38	0.61 / 0.63	2.58 / 2.10	0.10 / 0.12
100	0.00 / 0.00	0.78 / 0.75	0.00 / 0.00	0.12 / 0.11

Table 1: PFF/AUC$_f$ with topk=1, 10, 100.

6. Analysis of the Freeze Improvement.

Since our method gains better results in freeze issues, we hypothesize the improvement is brought by both the architecture design and mixed training method. We report the PFF in Table 1. In architecture, we sample tokens from the top-k tokens with the highest probability, instead of choosing the one with maximum probability as Bailando [9], which reduces the PFF. With extra HumanML3D data, the share motion decoder learns more motion sequence statics. Thus the PFF further improved. Thus both architecture and extra data mix training improve the PFF (AUC same).



(a) token-of-AIST++ (b) token-of-ml3d (c) token-of-share (d) val-loss

Figure 2: Shared tokens (latent space) with a human motion VQ-VAE architecture in Experiment B.

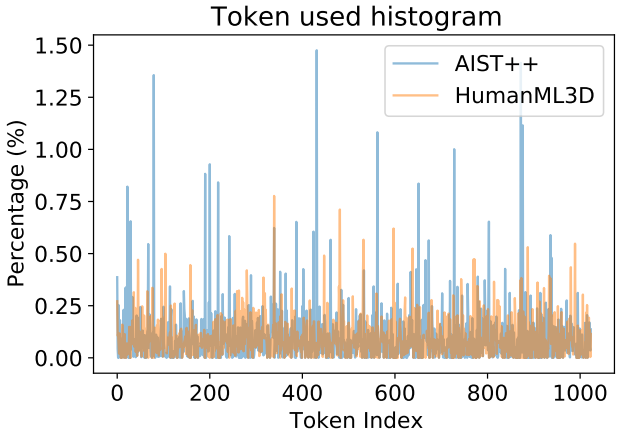


Figure 3: Token used histogram, histogram are normalized by the total frame from each dataset.

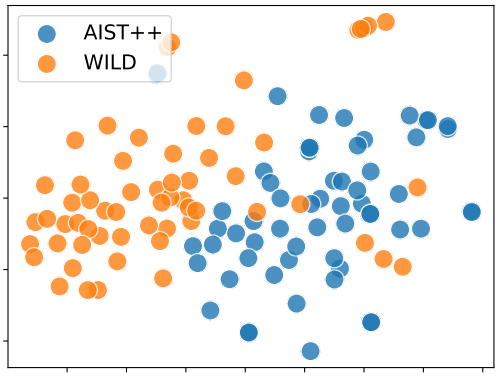
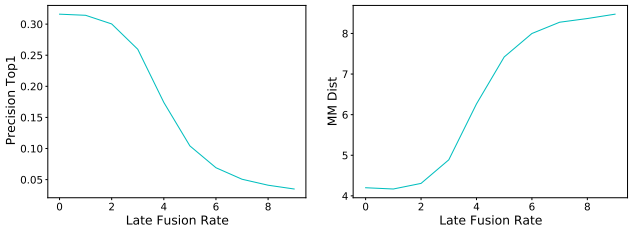


Figure 4: audio-t-SNE of datasets (orange: AIST++, blue: our dataset)).

7. More Visualizations of Our Results

We also show more visualizations of our results in the attached 'demo.mp4' file, which contains the following contents.



(a) top-1 (b) MM dist

Figure 5: The effect of LFR with t2m resulst.

- Comparisons with other music2dance methods in AIST++ test set and our in-the-wild dataset.
- Our results with the same music, different actions / time / durations.
- Comparisons with Slerp [8] for music-text conditioned dance generation.

From these videos, we can find that our results outperform other methods and are more realistic.

References

[1] Deepak Gopinath and Jungdam Won. fairmotion - tools to load, process and visualize motion capture data. Github, 2020. 1

[2] 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. 1, 2

[3] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision*, pages 580–597, 2022. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium.

In *Advances in neural information processing systems*, volume 30, 2017. 1

[5] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 2

[6] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *SIGGRAPH*, pages 677–685. 2005. 1

[7] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics*, pages 83–86, 2008. 1

[8] Ken Shoemake. Animating rotation with quaternion curves. In *Proceedings of the 12th annual conference on Computer graphics and interactive techniques*, pages 245–254, 1985. 3

[9] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 1, 2