OpenDance-3D: A Large-Scale Dataset for Music-driven Dance Generation

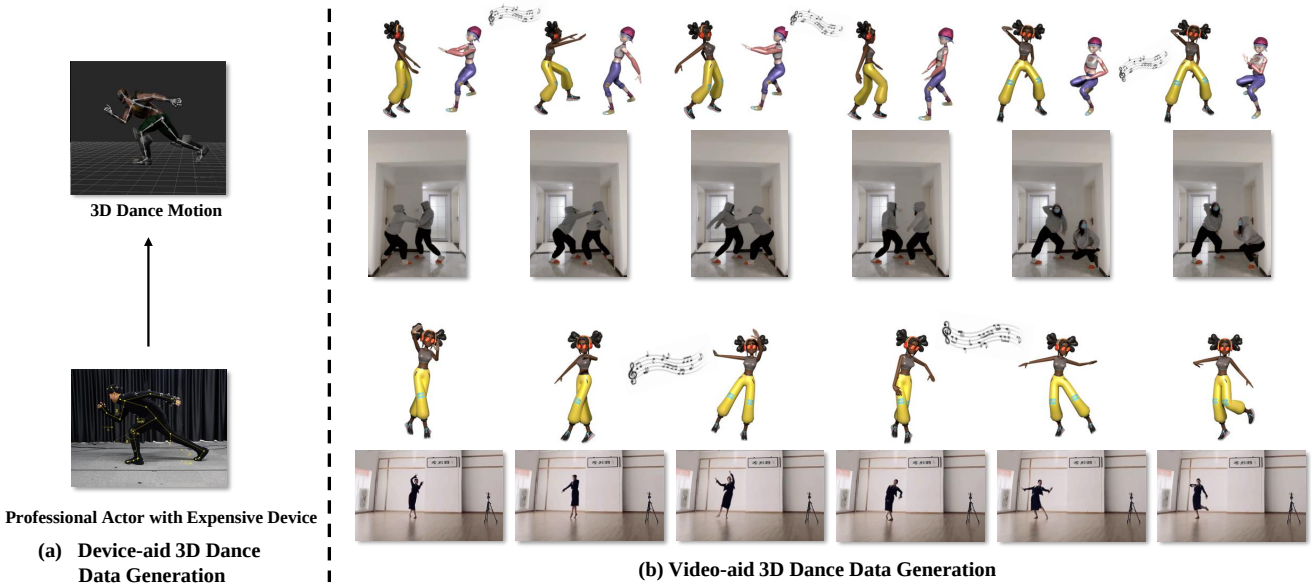

Figure 1. (a) Most existing datasets of 3D dance generation are heavily dependent on expensive devices and professional actors, and purely focus on solo dance and impede the development of large-scale 3D dance generation. (b) To break through the limitation of devices and actors for promoting 3D dance generation towards large-scale datasets, we subtly utilize 2D dance videos from the Internet to construct large-scale 3D dance motions. It is a cut-price way to obtain plentiful data without any devices and actors. By the way, we also involve collaborative dance (e.g., dance battle) in our dataset for the first time, which aims to explore the collaborative dancing of multiple people.

Abstract

Existing data collection of 3D dance generation is highly limited by expensive and professional devices and labor force, and is difficult to develop large-scale data of 3D dance motion for industry applications. To promote the development of large-scale 3D dance generation in the industry, we propose a new perspective to construct a large-scale dataset of 3D dance motion (i.e., ***OpenDance-3D****) by using 2D dance videos from the Internet rather than professional devices and actors. Our OpenDance-3D is featured from three aspects: 1)* ***Large scale****. It contains 28,051 3D music-dance sequence pairs with duration varying from 4.0 seconds to 120.0 seconds, resulting in the largest dataset in the wild to our knowledge. 2)* ***High diversity****. Created from Internet videos, it naturally consists of a wide range of dance motions and music genres that meet real-world user demands. 3)* ***Collaborative****. It contains 2,933 sequences to support collaborative dance motion generation (e.g., dance battle), which makes it possible for multiple people to dance in a 3D environment. To represent the relevance of collaborative dance motions, we carefully design a metric named motion-motion score. Besides, we propose a baseline called Bailando++ to handle collaborative dance motions. Extensive experiments show that our OpenDance-3D not only can better assist in model generalization to real-world data but also promote the development of large-scale pre-training. All data and codes will be made available upon acceptance.*

1. Introduction

Shall we dance? Shall we talk! —- *Eason Chan*

Music-guided 3D dance generation task [13, 15, 17, 25] has attracted more and more attention due to its huge potential applications, such as virtual idols (e.g., Hatsune Miku[1]), virtual choreographers for humans, and digital AI characters. However, most of the current approaches to produce music-conditioned 3D dance data (e.g., AIST++ [17]) are device-aid as shown in Figure 1 (a), which are extremely dependent on expensive and professional 3D motion-capture devices and actors. It obviously restricts the task to large-

[1]https://en.wikipedia.org/wiki/Hatsune_Miku

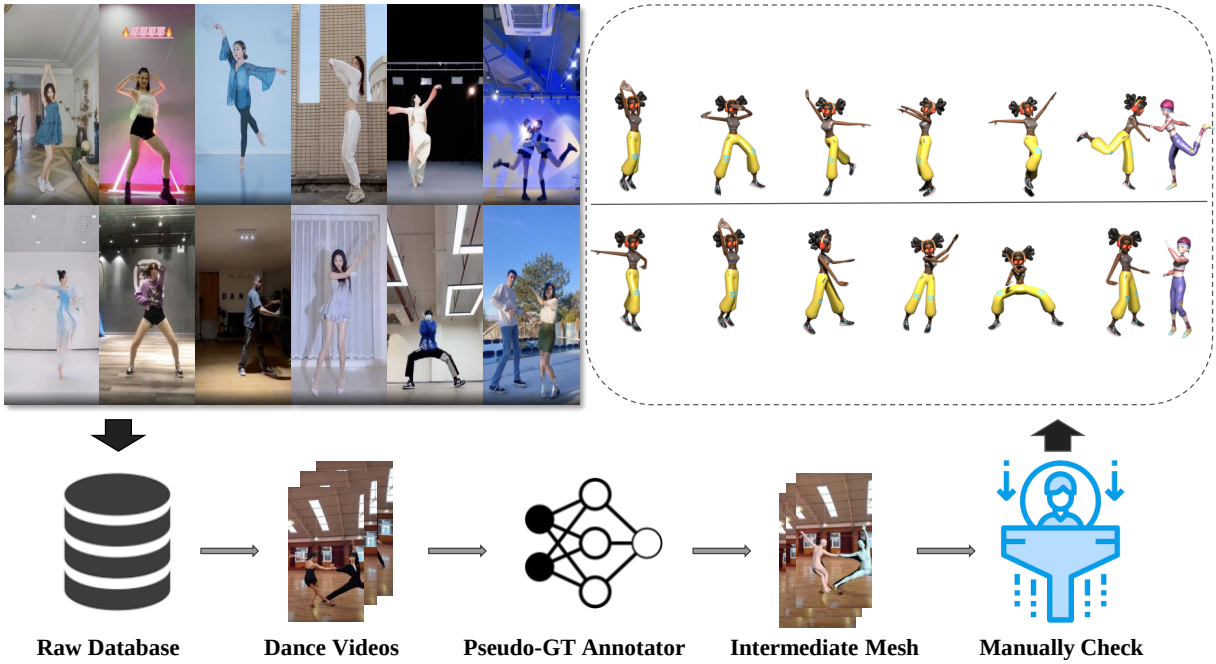

Figure 2. **Overview of the acquisition pipeline**. We show real-world dance videos (left) from our dataset and its corresponding animated avatars (right). Videos that fit specific criteria are downloaded, trimmed, pseudo-annotated, and undergo manual check.

scale learning and real-world generalization due to the idealistic environment of data acquisition. They also only cover limited dance and music genres by reason of the expensive labor force, which seriously impedes model generalization to real-world dance. Besides, Existing 3D dance datasets [2, 15–17, 26, 27, 34] only focus on solo dance. None of them consider the multi-person collaborative dance genres (e.g., dance battle and waltz, etc.). We believe the multi-person dance motion should be important in the music-driven 3D dance generation community and is also applicable to a wide range of real-world scenarios. In conclusion, there are three shortcomings for existing 3D choreography datasets: (1) limited scale and real-world generalization; (2) insufficient diversity of dance and music genres; (3) lack of collaborative dance genres.

To address the above shortcomings, we first discard device-aid data generation, and propose a novel way named video-aid 3D data generation as shown in Figure 1 (b) to build our large-scale dataset. Our video-aid 3D data generation borrows amounts of 2D dance videos from the Internet to build our 3D dance data. The whole pipeline of our video-aid 3D data generation can be seen in Figure 2. We call the dataset constructed from 2D video as **OpenDance-3D** since the data comes from the open world. Note that our OpenDance-3D is large-scale with 28,051 3D music-dance sequence pairs tailed with duration varying from 4 seconds to 120 seconds. To be more specific, we collect millions of dance videos from short video platforms with several filter strategies, and use advanced object detection [9], tracking [32], and 3D pose estimation [18] to recover the dance motion sequence. Towards the second shortcoming, we deliberately collect diverse music from various dance genres, including jazziness, ballet, hip-hop and folk dance, etc. Some of the them have lyrics, which contain English lyrics and Chinese lyrics. Note that there are too many music and dance genres, some of them we even do not know, so we are unable to count the number of them. Moreover, to the best of our knowledge, we are the first to propose multi-person collaborative dance (e.g., waltz, dance battle) for the music-conditioned 3D dance choreography task. To evaluate the generation quality of multi-person dance motion, we carefully design a metric named motion-motion score. At last, to improve the motion quality, we also apply motion completion for potential missing detection and occlusion and motion smoothing method to reduce frequent motion jitters, utilizing tools from MMHuman3D [7]. Furthermore, we employ a team of professional annotators to examine the quality of motion via visualization and drop out those in low quality, e.g., abnormal direction and extreme jitters.

Moreover, we evaluate the state-of-the-art method [25] on our OpenDance-3D and conduct ablation studies to analyze that our dataset not only can improve the real-world generalization of current methods but also demonstrate performance gains brought by pre-training on the large-scale video-aid data. To model the collaborative motions of multiple dancers, we present a novel baseline named Bailando++ because it is developed from Bailando [25].

The key merits of our paper lie in three aspects:

• We propose a large-scale 3D choreography dataset, using a video-aid 3D dance data generation to break through

the limitation of device-aid generation approaches, which makes the data style more realistic and diverse.

• To the best of our knowledge, we are the first to enrich current music-driven 3D dance generation task from solo dance to multi-person dance, which aims to explore 3D collaborative dance choreography. To model and evaluate multi-person dance, we propose a baseline called Bailando++ and design a motion-motion score concurrently.

• Sufficient experiments not only show qualitative results to assist the analysis of our OpenDance-3D characteristic (e.g., realistic and diverse choreographic) but also provide quantitative results to demonstrate the benefits of the dataset in real-world generalization and large-scale pretraining. Visualizations are in our supplementary materials.

2. Related Work

Music-conditioned 3D Dance Generation. Virtual AI choreographer which generates plausible dance motions conditioned on a piece of music, shows an especially promising prospect for its inherent entertainment properties. For example, virtual hologram music concerts that take place on short video platforms are becoming increasingly popular and cultivating many virtual idols. Different from the regular motion prediction task [1,4,8] that predicts future movements based on observed motion sequences, music-conditioned 3D dance generation poses more constraints on music rhythms and beat-wise motion coherence. Recent approaches [11,15,17,34] jointly consider both musical features and motion features in a space, and solve the task by regression processing. For instance, the FACT [17] employs an audio transformer and seed motion transformer to encode the inputs, and then conducts feature fusion using a cross-modal transformer to predict future motion sequences. The DanceFormer [15] further introduces two cascading kinematics-enhanced transformer-guided networks to tackle music-aligned long-term sequences with high kinematic complexity. Unlike direct joint feature embedding, a recent state-of-the-art method named Bailando [25] executes VQ-VAE [28] to encode and quantize dancing-style 3D pose. Then they use an actor-critic GPT [24] incorporated with the quantized dance codes and cross-conditional causal attention to generate next dance. Although these methods have continuously improved in dance quality and consistency between dance and music, due to the lack of relevant datasets, no work has yet considered the modeling of two-person interactive dance types. In our work, we further introduce this task into collaborative dance genre with a new baseline model called Bailando++.

Music-driven Dance Datasets. Producing realistic human dance motions has been long studied and has promoted many academic datasets, which can be roughly divided into two aspects: 2D and 3D. On the one hand, music-driven 2D dance datasets [11, 13, 14, 33] have easy access to capture large-scale data from the Internet. However, 2D dance data is not enough to meet the real demand of virtual AI choreographer, like Hatsune Miku. On the other hand, it is quite difficult to collect large-scale 3D dance data due to expensive devices and labor force. For example, [15] costs about 15 months to produce only 9 hours of dance animations. Most existing 3D dance datasets [2, 15, 17, 26, 27, 34] are device-aid to build the dataset in a laboratory environment. The device-aid way is not enough for large-scale applications in industry. Because of the laboratory environment, its diversity of music and dance genres is relatively constrained compared with real-world dances. Recently, some researchers [16] pay attention to such limitations and use 2D video from the Internet data to produce 3D dance data. Nevertheless, the average duration is only about 6 seconds per sample and cannot be used by other methods to generate long-term dance. They also neglect collaborative dance genre in their dataset. To solve above mentioned issues, we propose OpenDance-3D as the largest 3D dance dataset and the first to handle the collaborative dance genre with diverse music and dance motions.

3. OpenDance-3D Construction

3.1. Dataset Preparation

Dataset Collection. We acquire real-world paired dance-music 2D videos from short video platforms. In order to ensure the quality of dance, professional dancers are preferred. Thus, we first filter out active users who meet the following criteria: 1) The user has more than 100 followers. 2) The user has posted at least 2 videos in the past 90 days. 3) The user's post tag contains 'Entertainment-Dance' and its weight ratio is greater than 0.5. We found around 80,000 dance users, each posting an average of 15 videos. Finally, we obtain a total of over 1 million uncleaned dance videos.

Pre-Processing. Social media videos often also contain some other unrelated video clips, such as beginnings, endings, and special effects. Considering the complexity of videos in real scenes, we additionally utilize YOLOX [9] from MMDetection [5] and PySceneDetect[2] to filter multi-person scenes (group dance with more than 2 people usually repeats the same movement) and scenes without people. To separate each human motion, we adopt ByteTrack [32] from MMTracking [6] for tracking and Re-ID. Due to the characteristics of dance videos (usually don't have transitions), we hardly need to do additional cropping of single-dancer videos, and only re-locate the start and end times, so as to ensure that the video content always contains dance movements. All trimmed videos are converted to exact 60 FPS.

Pseudo-GT annotator. As it is hard to obtain the corresponding 3D ground truth of collected videos, so we ap-

[2]https://scenedetect.com/

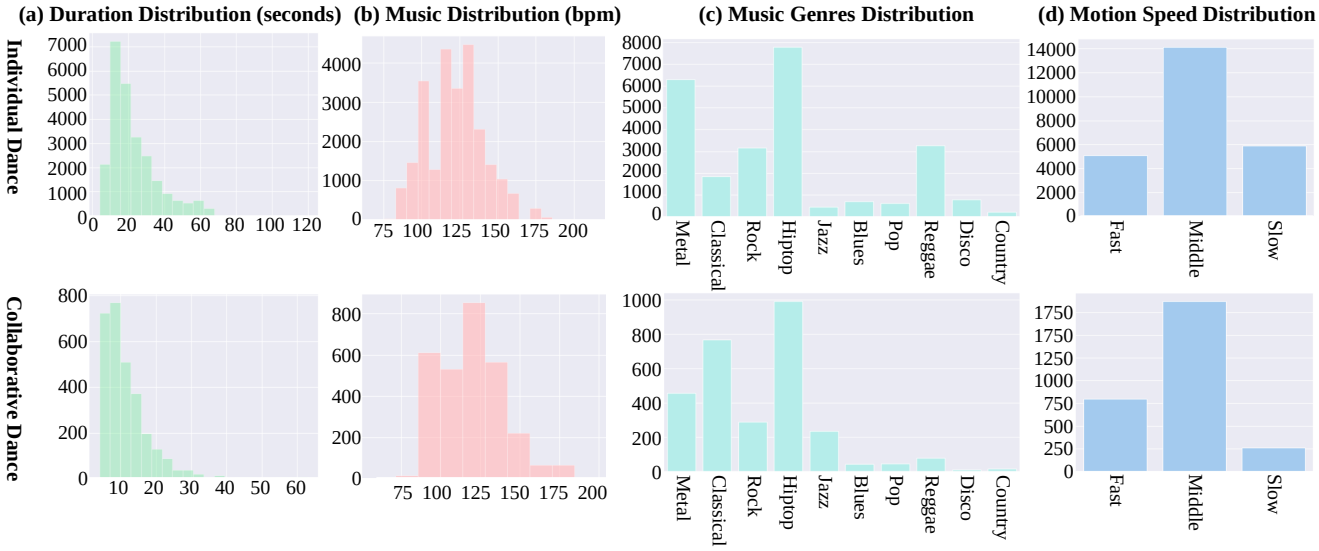

Figure 3. **Video number statistics on OpenDance-3D dataset including individual and collaborative dance genres.**

ply CLIFF [18], a state-of-the-art method for monocular, one-stage, regression of 3D human poses as our Pseudo-GT annotator. Afterward, we adopt linear interpolation and SmoothNet [31] from MMHuman3D [7] to handle missing detection from occlusion and frequent motion jitters. To further refine the results for better pixel alignment, we follow the same procedure as SPIN [12] and adopt SMPLify [3] fitting using the predicted results as initialization. We notice that there are still some abnormal results that model cannot handle, so we also introduce an additional manual inspection process to remove these bad results.

Manual Inspection. We invite more than 20 professional annotators to conduct 2 rounds of annotation. In the first round, the annotator will discard videos where: (1) the camera is moving dramatically; (2) there are more than 2 people or occlusions; (3) the dancer keeps his/her back to the camera all the time; (4) dance types with the help of other tools, such as pole dancing; (5) repeat the same movement or stand still. This round of cleaning is mainly to filter out dance videos that have quality problems themselves, and more than 90% of videos will be filtered at this stage. We then use CLIFF to generate meshes and 3D keypoints for the remaining videos in the first round of cleaning, and stitch the visualized results horizontally with the original video for the second round of cleaning. The second stage is mainly to filter the poor performance of the model, including abnormal orientation, severe jitter, and body jitter. The other filters are the same as in the first round to filter out the missing videos. After two rounds of cleaning, the final retention rate is around 2.5%.

Audio Processing. For audio, we follow the same audio processing as previous works [17, 25]. The music features are extracted by the public audio processing toolbox Librosa[3], including melfrequency cepstral coefficients (MFCC), MFCC delta, constant-Q chromagram, tempogram and onset strength, which are 438-dim in total.

3.2. Dataset Property and Analysis

The OpenDance-3D is a large-scale 3D human dance motion dataset that contains a wide variety of 3D motion paired with music in the wild. Each frame has the following annotations: 1) Pre-computed 24 SMPL [19] body pose joints in the kinematic tree. 2) Estimated SMPL pose, shape, and global translation. 3) Corresponding audio features. For the need of motion generation task, we release pre-extracted joints without raw videos to avoid potential copyright risk and obvious ethical considerations.

We perform statistical analysis on the trimmed dataset. OpenDance-3D provides 25,118 single-dancer videos and 2,933 collaborative dance videos without overlap. The distribution of video duration, music genres, music speed, and motion velocity is shown in Figure 3. The average sequence durations are 22.74 and 11.82 seconds, and the music speed (BPM) are in [68.0, 215.3] and [57.4, 198.8], for individual and collaborative genres respectively. For music genres, as we are unable to count the number of them, *we adopt the pre-trained model from [22] that classifies music into 10 classes (metal, classical, rock, hip-hop, jazz, blues, pop, reggae, disco, country) for reference.* Numerical values can be found in supplementary materials.

3.3. Comparison with Other Datasets

Table 1 shows the comparison between our OpenDance-3D dataset with the other music-conditioned dance datasets, in terms of the number of music pieces (Music), the number of dance genres (Genres), total duration (Hours) and etc.

[3]https://librosa.org/doc/0.8.1/

Table 1. **Comparison to existing music-dance datasets**. As far as we know, OpenDance-3D dataset is the largest public 3D dance datasets in the wild, and the first dataset that provides collaborative dance motions paired with diverse musics. 'AS' means average sample length (second). For device-aid methods, '*Mocap*' collects data via motion capture system. '*Multi-Cameras*' builds data from multi-views videos. '*Annotated*' means human annotation using industrial software. We also list several 2D relevant datasets for reference. The '/' means that is unmentioned in their paper. *Note that our dataset has too many music and dance genres that are unable to count.*

Dataset	Music	Genres	Hours	AS	2D/3D	Collaborative	Wildness	Public	Source
Listen2Dance [14]	/	1	6.26	/	2D	×	✓	✓	Internet
DancingToMusic [13]	/	3	71.00	5.0-10.0	2D	×	✓	✓	Internet
Dance Revolution [11]	/	3	12.00	/	2D	×	✓	✓	Internet
TikTok Dance-Music [33]	85	/	1.55	12.5	2D	×	✓	✓	Internet
DanceNet [34]	/	/	0.96	/	3D	×	×	×	Mocap
GrooveNet [2]	3	1	0.38	/	3D	×	×	✓	Mocap
EA-MUD [26]	23	6	0.51	/	3D	×	×	✓	Mocap
Dance with Melody [27]	61	4	1.57	/	3D	×	×	✓	Mocap
AIST++ [17]	60	10	5.19	13.3	3D	×	×	✓	Multi-Cameras
PhantomDance [15]	260	13	9.63	/	3D	×	✓	×	Annotated
YouTube-Dance3D [16]	/	/	70.00	6.0	3D	×	✓	×	Internet
OpenDance-3D (Ours)	/	/	168.38	22.74	3D	✓	✓	✓	Internet

Although 2D datasets [11, 33] exist, they are not enough to meet the real demand for 3D AI choreographers. While existing 3D datasets are mostly collected from optical motion capture systems [2, 26, 27, 34], reconstructed from multi-view videos [17], or annotated [15] using amounts of labor to facsimile dance motion in industrial animation software, which are expensive and time-costing. For example, [15] costs 15 months to finish the dance labeling of 9.63 hours of motion data. As a result, they are limited in scale and richness inevitably. AIST++ [17] only covers 60 pieces of music, and all of them are light music without lyrics. YouTube-Dance3D [16] is the most similar dataset to ours, but the average length of each clip is only about 6 seconds [15] which is too short to support long-term dance generation, in comparison with our 22.74 secs average length. In terms of collaborative dance genre, we are the first dataset that provides 3D collaborative dance motions. Even though [33] and [16] also have multi-person videos from the Internet, they are mostly group dances that repeat similar dance motions without deliberate collaboration, and each motion is handled individually in their procedures.

4. Approach

Problem Formulation. Given a seed motion sequence represented as $X = (x_1, ..., x_T)$ and a piece of music sequence represented as $Y = (y_1, ..., y_{T'})$, where $x_i \in R^{N*3}$, N=24 is the number of body joint. The target is to generate a sequence of future motion $X^{'} = (x_{T+1}, ..., x_{T'})$ from timestep $T+1$ to $T^{'}$ related to music style and rhythm, where the length of music $T^{'}$ is required to longer than T. Traditional task is usually presented as a conditional generation task as $P(X^{'}|X, Y)$ on paired music and motion. In this paper, we further involve with collaborative motion sequences and generalize it to $P(X_0^{'}, X_1^{'}|X_0, X_1, Y)$, where X_0 and X_1 are different but collaborative dance motions corresponding to a same music Y.

4.1. Individual Motion Generation

To handle individual motion generation, we adopt Bailando [25], the recent state-of-the-art method on this task as our primary baseline. Bailando involves two powerful components: 1) **Pose VQ-VAE**. It learns a choreographic memory (codebook) that summarizes meaningful dancing units from a series of 3D human motions in an unsupervised manner. 2) **Motion GPT**. It predicts the most possible dancing unit from the learned codebook given an initialized motion state and corresponding music. A fluent dance sequence is generated auto-regressively. The pipeline of Bailando is illustrated in Figure 4 (a).

4.2. Collaborative Motion Generation

Bailando++. Collaborative motion generation further poses new challenges on the coherence between motions. Thus, we build on the top of Bailando and develop a variant (Bailando++) to support multi-person motion generation conditioned on music. To be specific, we adopt a transformer layer with cross-conditional causal attention to explicitly enhance the relevance between motions. For comparison, we use raw Bailando which handles each human motion separately without building a relation between motions. The difference is shown in Figure 4 (b) and (c).

Motion-Motion Score. Inspired by music-motion beat

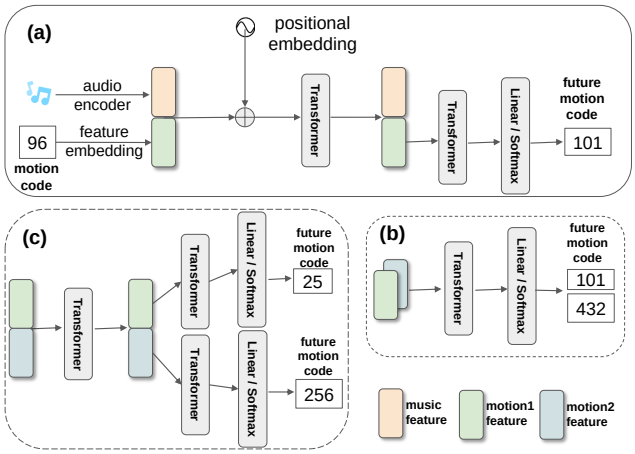

Figure 4. **Illustration of Bailando++**. (a) The pipeline of Bailando. (b) Handle each human motion separately (Bailando), conditional attention only works on music-motion. (c) Joint generation with interaction between moitions (Bailando++).

alignment score [17, 25], we define motion-motion score to evaluate relevance between motions. To illustrate, we first recall music-motion score, which calculates the average temporal distance between each music beat and its closest dance beat. The music beats are extracted using librosa and the motion beats are computed as the local minima of the kinetic velocity. Given two beat sequences of music B_{music} and corresponding motion B_{motion}, we take $f(B_{music}, B_{motion})$ as its alignment.

$$f(B_1, B_2) = \frac{1}{|B_1|} \sum_{t_1 \in B_1} exp(-\frac{\min_{t_2 \in B_2} \|t_1 - t_2\|^2}{2\sigma^2}), \tag{1}$$

where B records the time of beats, while σ is normalized parameter which is set to be 3 as default. However, unlike music-motion alignment where each music beat is asked to have a corresponding motion beat, the order (primary and secondary) of motions is unknown. Thus, we calculate it in a bi-directional manner as Equation 2.

$$\frac{1}{2}(f(B_{motion_1}, B_{motion_2}) + f(B_{motion_2}, B_{motion_1})). \tag{2}$$

In this case, motion-motion score well measures bi-directional alignment between B_{motion_1} and B_{motion_2}.

5. Experiments

5.1. Dataset and Baseline

Dataset. All the experiments are conducted on the OpenDance-3D dataset, which to our knowledge is the largest available dataset for this task. We split it into train and test sets as shown in Table 2, and report the performance on the test set only. For fine-tuning, we perform the training and evaluation on the AIST++ dataset. Large CPU memory (100GB at least) is required to load the full set.

Table 2. **Data Splits based on Music-Choreography.**

	Type	Train	Val	Test
Hours	Individual	142.84	7.96	7.86
	Collaborative	8.93	0.36	0.33
Sequences	Individual	22606	1256	1256
	Collaborative	2733	100	100

Individual Baseline. The music-driven 3D human motion generation is an emerging task and we adopt the recently proposed state-of-the-art method, Bailando as described in Sec 4.1. We use the default setting using the official code provided by the authors. We analyze how the dataset scale matters for model performance by splitting the training set into several sub-splits. Specifically, to decrease training time, we train Pose VQ-VAE for 150 epochs, while for Motion-GPT, it takes more time to converge on a large dataset, and we train 100/200/400/600 epochs for $1K$, $5K$, $15K$ and $Full$ training set, respectively. More adequate training and refined config may lead to improvement.

Collaborative Baseline. For collaborative motion generation, we construct Bailando++ on top of the individual baseline, which is learned to generate correlated motions conditioned on the same music. The design of Bailando++ is described in Section 4.2. For comparison, we take Bailando that trained on individual motion-music pairs for reference, which handles each motion separately without any consideration for motion-motion correlation.

5.2. Evaluation Metrics

We inherit criteria from prior works [17, 25] and evaluate the generated results on the following aspects: (1) motion quality (2) motion diversity (3) motion-music correlation (4) motion-motion correlation. Specifically, for motion quality, we calculate the Frechet Inception Distances (FID) [10] between the generated dance and all motion sequences (including training and test data) of the OpenDance-3D dataset on kinetic features [21] (denoted as k) and geometric features [20] (denoted as g). For motion diversity, we compute the average feature distance of generated movements as [17]. As for the beat correlation, we calculate the corresponding motion-music score and motion-motion score as defined in Equations 1 and 2, respectively. Motion-motion score is only considered for collaborative generation tasks.

5.3. Evaluation Results

We conduct experiments for individual motion generation and collaborative motion generation on corresponding datasets, respectively. First, we conduct ablation studies on dataset scale for individual motion generation task,

Table 3. **Quantitative results on OpenDance-3D test set** for individual and collaborative motion generation respectively.

	Method	$FID_k \downarrow$	$FID_g \downarrow$	$Div_k \uparrow$	$Div_g \uparrow$	Motion-Music Score↑	Motion-Motion Score↑
Single	Ground Truth	-	-	3.86	7.21	0.2424	-
	Bailando$_{1K}$	54.23	20.54	2.36	4.83	0.2392	-
	Bailando$_{5K}$	49.86	17.77	2.32	5.15	0.2486	-
	Bailando$_{15K}$	47.01	13.06	2.79	5.86	0.2485	-
	Bailando$_{Full}$	46.65	12.84	2.82	6.18	0.2497	-
Multiple	Ground Truth	-	-	7.50	7.22	0.2352	0.3109
	Bailando	68.81	33.97	2.97	5.89	0.2562	0.2678
	Bailando++	65.13	32.19	2.96	5.94	0.2598	0.2748

and show that larger dataset scale matters for model performance. We sample sub-dataset into $1K$, $10K$, $15K$ and $Full$ incrementally and evaluate on the same test set. The quantitative results are shown in Table 3 (Single). As expected, the model performance consistently improves on motion quality, diversity and its alignment with music. We also find that the gain decreases gradually as the dataset size increases. For example, the model trained on $15K$ and $Full$ set achieve very close results. We attribute this to the mechanism of Bailando, where the learned codebook (choreographic memory) tends to get saturated gradually as the dataset size increases.

For the new proposed collaborative motion generation task, we take the original Bailando that handles each motion separately for comparison, and evaluate its extension Bailando++ under the same setting. We show the results in Table 3 (Multiple). Enhanced by motion-motion interaction module as designed in Figure 4 (c), the motion-motion score achieves better in comparison with individual generation as Figure 4 (b), along with gains on other metrics. But slightly unexpectedly, Bailando also achieves 0.2678 on motion-motion score. We conjecture that individual interaction between motion and music has already implicitly closed up the gap between motion and motion, as the music is shared between motions.

Rethinking about Music-Motion Score. We find that the metric of the generated dances is even higher than the ground truth on the correlation between motion and music. Although we have conducted manual cleaning, there may still exist potential unexpected cases in the dataset. But more importantly, it motivates us to rethink the definition of music-motion score. It starts from each music beat, the basic rhythmic unit of a measure, the interval between two adjacent beats is usually unchanged within music. However, the periodical music beat is not necessarily identical to the underlying discrete pulse of a piece of music. This raises two issues. First, quantitative results can diverge from subjective results. Besides, due to the periodic of music beats, the metric is naturally friendly to dense motion beats, no matter what kind of music is given. Figure 5 gives an example, where denser velocity minima (most likely from motion jitter) in generated dance (upper) lead to an inflated alignment score in comparison with ground-truth (lower). This remains an unresolved issue for future works.

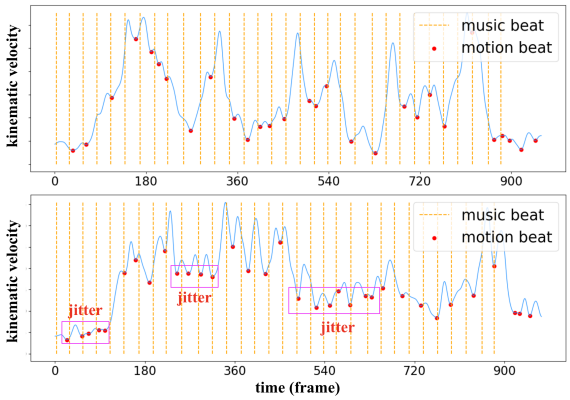

Figure 5. **Illustration of motion-music score**. Motion beats (red dot) correspond to local minimal of kinematic velocity (blue curve). Orange vertical line represents positions of music beats. The ground-truth (upper) and generated movement (lower) achieve 0.2533 and 0.3199 score respectively. Obviously, denser motion beats (usually from motion jitter) lead to inflated evaluation score.

5.4. Generalization Performance

Real-world Generalization. To demonstrate the benefits of a large-scale dataset in real-world generalization, we select 25 pieces of music in the wild and ensure they do not existed in the dataset. For a more intuitive understanding of the quality of the generated dance, we conduct a user study to subjectively score the dance sequences generated by Bailando trained on OpenDance-3D and AIST++ respectively. No fine-tuning is adopted. We invite 10 participants that are experienced with subjective evaluations. For each participant, we randomly play 25 pairs of comparison videos (the order is unknown) with a length of 10 to 25 seconds, where each pair contains our result and one competitor's in the same music. We ask the participant 3 questions to indicate "is the dance coherent in time?", "is the movement diverse?" and "is the dance movement relevant to the music beat?", correspond to the 'music-motion score', 'motion diversity' and 'motion quality'. For each question, they are asked to score each dance from 5 choices [0, 0.25, 0.5, 0.75,

1]. The statistics are shown in Table 4.

Table 4. **Generalization performance on unseen music.** We report the average score over three subjective metrics, where *coherence* shows the quality of movement temporally, *diversity* shows the overall number of different movements, *correlation* shows the relevance of motion-music.

Dataset	Coherence	Diversity	Correlation
AIST++	**0.5814**	0.5114	0.4429
OpenDance-3D	0.5571	**0.5614**	**0.5729**

As expected, our dance motions show better diversity and correlation, which means that the model trained on a large-scale dataset (OpenDance-3D) is equipped with better generalization performance, the generated dance motions are more plausible given any unseen music. However, we find that on coherence we are slightly worse than the model trained on AIST++, which is consistent with the gap in absolute quantitative results of motion quality. We attribute this to the gap between estimated motion and captured motion in quality. We show real visualization results (both good and bad) in supplementary materials.

Analysis of Choreographic Memory. To demonstrate our OpenDance-3D can bring better diverse dance motions with real-world generalization compared with the device-aid dataset, we train Bailando++ on different datasets (OpenDance-3D vs AIST++). Then we visualize the latent codes from the codebook in 2D space via t-SNE [29], which implicitly assesses the diversity of motion in each dataset.

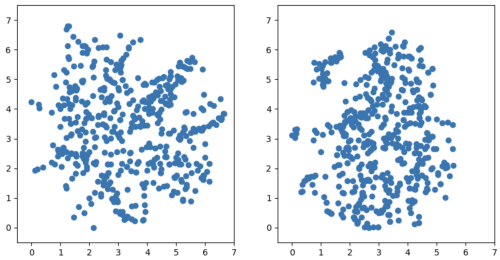

Figure 6. **t-SNE visualization** of 512 dance units from choreographic memory trained on OpenDance-3D (left) and AIST++ (right). The distribution indicates the diversity of codebook.

As shown in Figure 6, the dance units trained on OpenDance3D are more sparsely distributed than AIST++, as the capacity of choreographic memory trained on OpenDance-3D improves. It indirectly elucidates that open-world dances indeed cover more diverse dance motions.

5.5. Pre-training Matters

Large-scale datasets in the wild have been validated usefulness for pre-training in cross-modality tasks [23, 30]. To verify it on our task, we pre-train the VQ-VAE on the full training set of OpenDance-3D (individual) for 100 epochs and then finetune it on AIST++. All other following steps are kept in the same setting as Bailando without meticulous adjustment. Pre-training with more epochs may lead to better performance. We report the results of VQ-VAE and the final results in Table 5 and Table 6, respectively.

Table 5. **Ablation study on AIST++ test set.** The pre-training is conducted on OpenDance-3D$_{train-full}$ first before finetuning on AIST++. Experiments are conducted on pose VQ-VAE, which indicates the quality of learned dance units.

Method	FID$_k$ ↓	FID$_g$ ↓	Div$_k$ ↑	Div$_g$ ↑
w/o pre-trained	28.23	12.63	6.80	6.57
w pre-trained	26.81	12.06	7.13	6.87

Table 6. **Quantitative results on AIST++ test set.** FACT and Bailando are multiplexing the same results of AIST++ benchmark . * is the flag for pre-training.

Method	FID$_k$ ↓	FID$_g$ ↓	Div$_k$ ↑	Div$_g$ ↑	BAS↑
FACT	35.35	22.11	5.94	6.18	0.2209
Bailando	28.16	9.62	7.83	6.34	0.2332
Bailando*	26.22	8.72	7.91	6.71	0.2355

As shown, pre-trained on our dataset, a diverse choreographic memory with higher quality is learned to summarize meaningful dancing units. We believe that a large-scale dataset can ease the degree of over-fitting and improve the diversity of generated motions, and be helpful for other data-hungry model structures and the whole community.

6. Conclusion and Discussion

We propose a video-aid data generation way to build a large-scale music-driven 3D dance dataset named OpenDance-3D, which consists of single-person and multi-personal collaborative dance motions with diverse music and dance genres. To explore the collaborative dance, we propose a method named Bailando++ with a newly designed metric called motion-motion score. Sufficient experiments show that our OpenDance-3D not only can better assist in model generalization to real-world data but also promote the development of large-scale pre-training.

Discussion. There are many interesting research directions that can be grown up in this paper. For example, it is under-explored for more than three persons to interactively dance. How to build a more elegant network for collaborative dance is still an exciting topic in our future work.

References

[1] Vida Adeli, Mahsa Ehsanpour, Ian Reid, Juan Carlos Niebles, Silvio Savarese, Ehsan Adeli, and Hamid Rezatofighi. Tripod: Human trajectory and pose dynamics forecasting in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13390–13400, 2021. 3

[2] Omid Alemi, Jules Françoise, and Philippe Pasquier. Groovenet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17):26, 2017. 2, 3, 5

[3] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pages 561–578. Springer, 2016. 4

[4] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6158–6166, 2017. 3

[5] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 3

[6] MMTracking Contributors. MMTracking: OpenMMLab video perception toolbox and benchmark. https://github.com/open-mmlab/mmtracking, 2020. 3

[7] MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark. https://github.com/open-mmlab/mmhuman3d, 2021. 2, 4

[8] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6519–6527, 2020. 3

[9] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 2, 3

[10] 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. *Advances in neural information processing systems*, 30, 2017. 6

[11] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 3, 5

[12] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 4

[13] 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, 3, 5

[14] Juheon Lee, Seohyun Kim, and Kyogu Lee. Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network. *arXiv preprint arXiv:1811.00818*, 2018. 3, 5

[15] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1272–1279, 2022. 1, 2, 3, 5

[16] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171*, 2020. 2, 3, 5

[17] 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, 3, 4, 5, 6

[18] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *ECCV*, 2022. 2, 4

[19] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 4

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

[21] Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. Fmdistance: A fast and effective distance function for motion capture data. In *Eurographics (Short Papers)*, pages 83–86, 2008. 6

[22] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. Rethinking cnn models for audio classification. *arXiv preprint arXiv:2007.11154*, 2020. 4

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 8

[24] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3

[25] 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, 3, 4, 5, 6

[26] Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S Kankanhalli, Weidong Geng, and Xiangdong Li. Deepdance: music-to-dance motion choreography with adversarial learning. *IEEE Transactions on Multimedia*, 23:497–509, 2020. 2, 3, 5

[27] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1598–1606, 2018. 2, 3, 5

[28] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *NIPS*, 2017. 3

[29] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 8

[30] Jue Wang, Haofan Wang, Jincan Deng, Weijia Wu, and Debing Zhang. Efficientclip: Efficient cross-modal pre-training by ensemble confident learning and language modeling. *arXiv preprint arXiv:2109.04699*, 2021. 8

[31] Ailing Zeng, Lei Yang, Xuan Ju, Jiefeng Li, Jianyi Wang, and Qiang Xu. Smoothnet: A plug-and-play network for refining human poses in videos. *arXiv preprint arXiv:2112.13715*, 2021. 4

[32] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 2, 3

[33] Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. Quantized gan for complex music generation from dance videos. *arXiv preprint arXiv:2204.00604*, 2022. 3, 5

[34] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022. 2, 3, 5