# FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation

Ronghui Li[1*], Junfan Zhao[1*], Yachao Zhang[1],
Mingyang Su[1], Zeping Ren[1], Han Zhang[2], Yansong Tang[1], Xiu Li[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Northwestern Polytechnical University
{lrh22, yf-zhao21}@mails.tsinghua.edu.cn, {yachaozhang, tang.yansong, li.xiu}@sz.tsinghua.edu.cn

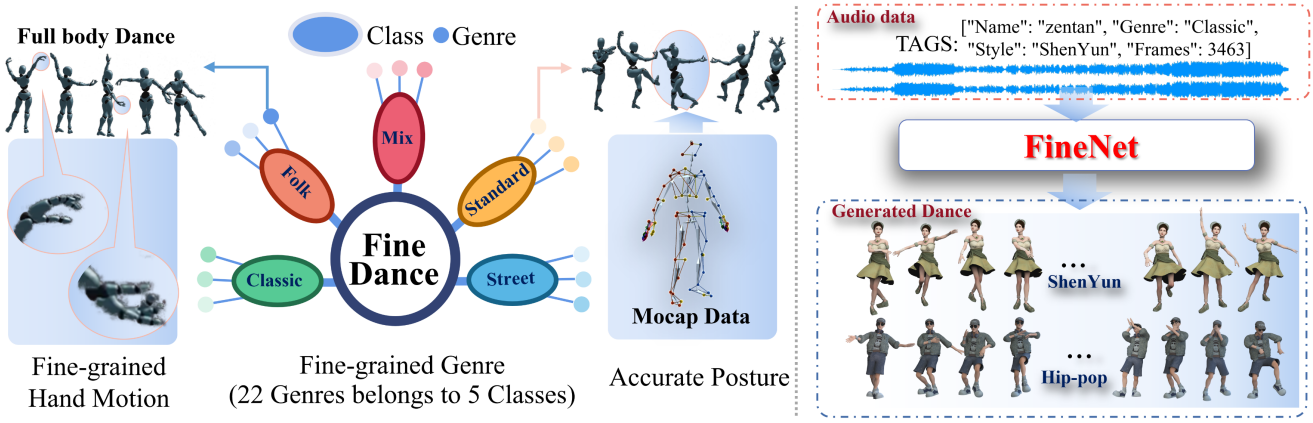


Figure 1. A conceptual overview. We release a large-scale professional 3D motion capture dance dataset **FineDance**, and propose a choreography network **FineNet**. Training with FineDance, FineNet can generate multi-genre dances with expressive hand movements.

## Abstract

*Generating full-body and multi-genre dance sequences from given music is a challenging task, due to the limitations of existing datasets and the inherent complexity of the fine-grained hand motion and dance genres. To address these problems, we propose FineDance, which contains 14.6 hours of music-dance paired data, with fine-grained hand motions, fine-grained genres (22 dance genres), and accurate posture. To the best of our knowledge, FineDance is the largest music-dance paired dataset with the most dance genres. Additionally, to address monotonous and unnatural hand movements existing in previous methods, we propose a full-body dance generation network, which utilizes the diverse generation capabilities of the diffusion model to solve monotonous problems, and use expert nets to solve unreal problems. To further enhance the genre-matching and long-term stability of generated dances, we propose a Genre&Coherent aware Retrieval Module. Besides, we propose a novel metric named Genre Matching Score to evaluate the genre-matching degree between dance and music. Quantitative and qualitative experiments demonstrate the quality of FineDance, and the state-of-the-art performance of FineNet. The FineDance Dataset and more qualitative samples can be found at website.*

## 1. Introduction

Music and dance are two enduring art forms that can express a wide range of human emotions, and they have become essential elements in modern entertainment industries such as concerts, movies, and games[11]. However, creating high-quality 3D dance animations can be a costly and complex process that often involves skilled dancers, engineers, and expensive motion-capture equipment[5]. As a result, there is growing interest in using artificial intelligence (AI) to generate 3D dance animations from music, which

∗ equal contribution, † corresponding author

has become a rapidly developing research topic.

Despite the wide range of research in this field [19, 14, 22, 21, 2, 49, 35, 5, 44, 36, 25, 46, 47], enerating high-quality dances is still limited by the existing datasets: (1) **Full-body expressiveness:** Existing dance datasets contain few hand movements, and only 1.5 hours of full body dance data is available. The previous methods in motion generation, treating body and hand as the same, lead to unnatural or monotonous hand motions, because the body and hand are in different feature spaces. However, uncoordinated body and hand motions can destroy the expressiveness of the overall dance. (2) **Multi-genre:** Existing datasets contain a limited number of dance genres, so the generated dances are not sufficient to match various music styles. Previous methods struggle with limited coarse dance genres and have no suitable objective metric to measure the genre-matching degree between music and generated dances.

To address the limitations of existing datasets, we introduce a Fine-grained Choreography Dance dataset (FineDance). It comprises over 14.6 hours of data collected from 346 paired songs and dances, was created by professional dancers and a motion capture system, which has accurate body and hand motions. The fine-grained 22 dance genres of FineDance spanning traditional and modern styles, which make the genre-matching of generated dance sequences and given music become more challenging. FineDance includes music, dance sequences, FilmBox (fbx) files, SMPL[23, 29], and multi-view videos.

Early music-driven dance synthesis methods [3, 16, 24, 32, 15, 35, 2, 48] often rely on motion graph-based algorithms where the dance fragments from a pre-existing music-dance database are stitched together to synthesize one dance. While such methods can synthesize long-term dances, they do not produce new dance fragments. Recently methods have employed generative networks such as VAE[33], GAN[30], Normalization Flow Network[40], Diffusion[38]. But they focus solely on body part, while neglecting hand movements, resulting in unnatural or monotonous hand motions even trained with well-annotated body and hand labels. Additionally, the generative-based methods are limited by the long-term modeling ability of the networks, making them difficult to generate long-term dance sequences.

Therefore, we propose FineNet, a two-stage generative-synthesis network that addresses the limitations of previous dance generation methods. In the first stage, we propose a diffusion-based Full-body dance generation network (FDGN). The key of FDGN is to design two expert networks, which are dedicated to the generation of body and hand motions, and use a Refine Net to assemble them coordinately. In the second stage, we propose a Genre&Coherence aware Retrieval Module (GCRM), which ensures the coherence of dance fragments and matches the genre between the music and the dances. Based on the suitable dance fragments retrieved by GCRM, we can produce genre-matching and long-term dances. A conceptual overview of the dataset and method is shown in Figure1. Finally, to objectively evaluate the genre-matching degree between generated dances and given music, we propose a novel metric, named Genre-matching Score (GS).

Overall, our contributions can be summarized as follows:

- We release FineDance, which is the largest 3D motion capture music-dance paired dataset with accurate full-body posture, containing 22 fine-grained genres. FineDance encourages the development of AI choreography, motion prior, and full-body reconstruction methods.

- We present FineNet, which leverages expert networks and refine network to generate expressive full-body dances, and employs a cross-modal retrieval network to improve genre-matching scores.

- Extensive quantitative experiments and user studies demonstrate that our approach can generate multiple different genre-matched dances from arbitrary music with natural and flexible hand movements.

## 2. Related Works

### 2.1. Choreography Dataset

Currently, the most popular 3D choreography dataset is AIST++ [22], which provides 5 hours of dance but does not have hand motions. AIST++ is reconstructed through multi-view video. Therefore, there is inevitably a deviation between the generated 3D dance motions and the real motions because of the reconstruction error. Li *et al.* [21] provided another major type of 3D dance dataset: modeling in software, which is obtained by experienced animators. However, this type of dataset lacks the authenticity of the dance. At present, the most accurate datasets are obtained from motion capture systems. GrooveNet dataset [2] only contains the electronic dance genre in 23 minutes sequence. Dance with Melody [35] further constructed a 94-minute 3D dance dataset with 4 genres. Music2Dance [49] is a dataset with only one hour and 2 genres (modern and curtilage dance). Chen *et al.* [5] built a 9 hours dataset, but it only has four different dance genres. This genre partition does not match the discernment of professional dance artists. For example, Anime and K-pop do not count as specific dance genres. In conclusion, these datasets have limitations in hand motion, duration, and the diversity of genres and dancers. Our dataset contains 14.6 hours with 22 genres, and is collected by 27 dancers.

## 2.2. AI Choreography

**Synthesis based approach.** Early works[28, 26, 4, 43, 20] usually synthesis dances based on motion graphs and databases, which share one core idea: retrieving the most matching dance fragment for the given music clip, and splicing multiple fragments into a complete dance. In computer graphics, motion synthesis has long attracted a lot of attention as it can synthesize 3D actions from existing motion databases. Lamouret *et al.* [18] first delivered this idea and proposed a prototype system to create new actions by cutting and pasting action fragments from an action database. Arikan *et al.* [3] formally introduced the concept of graph-based motion synthesis, transforming this problem into finding paths in a pre-built motion graph. Under such a framework, the task of synthesizing actions is usually regarded as finding the optimal path in a constructed motion graph. Similar utilizing this graph-based scheme, Kim *et al.* [16] made the first attempt to synthesize rhythmic movements by adding constraints connecting action beats and rhythmic patterns. Shiratori *et al.* [32] and Kim *et al.* [15] formally pointed the music-driven dance movement synthesis problem and further developed more complex rules to associate dance movement segments with input music segments. This type of method can synthesize dances that match the music style well. However, as it is unable to learn the internal connections between music and dance, the synthesized dances cannot match the rhythm well. In addition, because the dance fragments are all from the database, such methods have no ability to create new dance motions. In this paper, we introduce the diversity generated technology into the synthesis method to maintain the advantages of synthesis and avoid the above problems.

**Generation based approach.** Generative Adversarial Network (GAN) [10] and Variational Auto Encoder (VAE) [17] have been successfully applied to generate various data modalities, including image, motion, music, etc. As a result, researchers also propose music-driven dance generation algorithms based on the general deep generation paradigm[7, 22, 34, 1, 21, 8, 42, 13, 12]. Such methods can be divided into 2D and 3D solutions according to the dimension of the generated dance data. Lee *et al.* [19] proposed the first music-driven 2D dance generative network, which uses VAE to model dance units and GAN to loop to generate dance sequences. Since the human skeleton is a natural graph data, Ren *et al.* [30] and Ferreira *et al.* [6] adopted a Graph Convolutional Network to improve the spatial naturalness of the generated 2D dance movements. Both music and dance belong to sequence data. Therefore, Li *et al.* [22] used the Transformer [41] network with strong sequence modeling ability to design a music-driven 3D dance action generation network. Recently, diffusion-based networks succeed in text2motion generation[37, 45]. Although these generative networks have the advantages, such as rhythm-matched and diversity, they neglect the quality of generated hand motions, and only generate motion fragments within several seconds. Our method can generate expressive full-body dances due to our diffusion based expert nets, and use a retrieval based modal to enhance the genre-matching score and take the advantages of synthetic methods such as genre-matching, long-term stability.

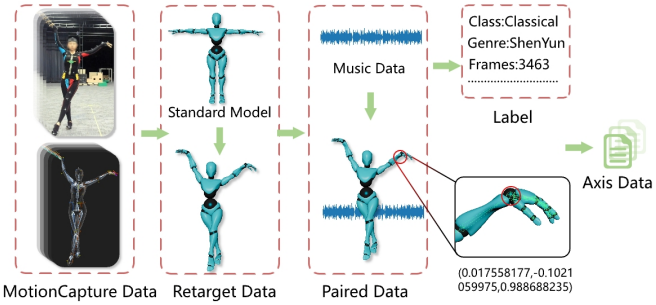


Figure 2. FineDance acquisition process. We capture the dancers' motions with the Vicon optical motion capture system. Then engineers retarget it to a standard skeleton in MotionBuilder. Dancers manually align the music with the dance motions and extract the skeletal information of the dance in Blender.

# 3. FineDance Dataset

The review of the choreography dataset is shown in Section 2.1, there are still few accessible large-scale motion capture choreography datasets even though many papers reported their choreography datasets. Meanwhile, most of these datasets are insufficient to train a diversified and generalized-well choreography model due to the limiting of dance genres, and poor music-dance pairings. Our FineDance can fill these gaps.

## 3.1. Data Acquisition and Analysis

For the fine-grained dataset, we take the following regulations into account for our dataset acquisition. We give the flowchart of the process shown in Figure 2. We summarized the comparison of FineDance and existing 3D dance datasets in Table 1.

**Fine-grained motions.** So far, there are only 2 datasets containing finger motion, and the available data is less than 1.5 hours. Fine-grained motions are ignored in existing methods. For example, GrooveNet[2] only contains 30 body joints, and Dance with Melody[35] has only 21 body joints, while EA-MUD[34], PhantomDance[21] and AIST++[22] all use 24 body joints of the SMPL[23] model.

Our data store the information of the skeleton joints in 3D space in each frame including fingers, which can help to improve the artistry and reality of the dance motion. For easy to utilize, we use the standard 52 joints to represent the dance data.

**Fine-grained dance genres.** Previous literature focuses on a few dance genres, such as GrooveNet[2],

| Dataset | Pos/Rot | Joints num | Hand joint | Genres | Mocap | RGB Views | Fbx | SMPL | Dancers | Total hours | avg Sec per Seq |
|---|---|---|---|---|---|---|---|---|---|---|---|
| GrooveNet[2] | ✓/✗ | 30 | ✗ | 1 | ✓ | 1 | ✗ | ✗ | 1 | 0.38 | **690** |
| Dance w/Melody[35] | ✓/✗ | 21 | ✗ | 4 | ✓ | 4 | ✗ | ✗ | - | 1.6 | 92.5 |
| Music2Dance[49] | ✓/✗ | 55 | ✓ | 2 | ✓ | 2 | ✗ | ✗ | 2 | 0.96 | - |
| EA-MUD[34] | ✓/✓ | 24 | ✗ | 4 | ✗ | 4 | ✗ | ✗ | - | 0.35 | 73.8 |
| PhantomDance[21] | ✓/✗ | 24 | ✗ | 13 | ✗ | 0 | ✗ | ✗ | - | 9.6 | 133.3 |
| AIST++[22] | ✓/✓ | 17/24 | ✗ | 10 | ✗ | **10** | ✗ | ✓ | - | 5.2 | 13.3 |
| MMD[5] | ✓/✓ | **52** | ✓ | 4 | ✓ | 0 | ✓ | ✗ | - | 9.9 | - |
| **FineDance (Ours)** | ✓/✓ | **52** | ✓ | **22** | ✓ | 2 | ✓ | ✓ | 27 | **14.6** | 152.3 |

Table 1. Comparisons of 3D Dance Datasets. Pos and Ros means 3D position and Rotation information respectively. Fbx (FilmBox) is one of the main 3D exchange formats as used by many 3D tools. "avg Sec per Seq" means the average seconds per sequence.

Music2Dance[49], MMD[5] and Dance with Melody[35]. This dataset uses a rough genre classification strategy, which is a non-standardized dance division.

We improve the diversity of our dataset from two aspects: more genres and more dancers. Our FineDance is reasonably classified under the advice of dance artists, covering hip-hop and Chinese classical dance more completely. To the best of our knowledge, it also includes folk dance motions for the first time, expanding the dance genres of the choreography dataset. Totally, FineDance has up to 22 genres of dance defined by professional dance artists. And we obtained more than 14 hours of data. It is worth noting that FineDance contains the most genres. Details are given in the supplementary materials.

**Accurate posture.** Currently, the largest available dataset is AIST++[22], which is collected by reconstructing 3D poses in multi-view videos. But the dance data is not real due to the reconstructing errors. Instead of reconstruction from the videos, ours is collected by a motion capture system, and all dance motions and music are well paired.

In FineDance, all motions are captured by the Vicon optical motion capture system and retargeted to a standard skeleton in MotionBuilder by engineers. Therefore, FineDance can donate accurate postures. Moreover, FineDance will be the largest fully available 3D music-dance paired dataset, and it will be available.

**Well-paired dance and music.** Dance fragments are strongly associated with the rhythm and style of music. However, due to the lack of enough well-paired data, the generative model is hard to fit the relevance of the motion rhythm and music rhythm. Therefore, we asked the professional dancers to pay attention to the matching of rhythm and style when dancing.

**Professional dancer.** We invited 27 professional dancers, and each dancer was asked to dance to the music while his/her motions were captured utilizing the capture system.

# 4. FineNet

## 4.1. Overall Framework

Given music of unknown style and arbitrary duration, our goal is to generate multiple different full-body dances. This task presents challenges in three areas: (1) Full-body: Body and two hands motions in different spaces have different grains, using a single network to generate full-body motions like the previous method can lead to unreal and monotonous hand gestures. (2) Genre-matching: Making the generated dances consistent with the fine-grained genre is also challenging due to the modal gap between music and dance. (3) Long-term: Generating long-term novel motion is challenging because neural networks tend to accumulate errors over time. To address these issues, we propose FineNet, which comprises a Diffusion-based Full-body Diverse Dance Generation Network (FDGN) and a Genre & Coherent aware Retrieval Module (GCRM). The FDGN focuses on creating detailed dances with expressive movements, while the GCRM considers the overall choreography of the dance. FineNet cleverly combines generative and synthetic methods, making them complementary, much like the process of human choreography. Furthermore, FineNet allows for the generation of multiple different dances by selecting distinct dance segments at the initial time step. This capability offers users a wide range of creative possibilities.

The overall framework of our method is shown in Figure 3. First, the input music $X$ is split into 4-second clips $\{X^t\}_{t=1}^N$ without overlapping, and $N$ is the number of clips of the given music.For each $X^t$, we use the Librosa toolbox [27] to extract the temporal feature $\bar{X}^t \in \mathbb{R}^{T\times C^m}$, and the mel-spectrogram image $\widetilde{X}^t \in \mathbb{R}^{W\times H\times 3}$, where $T$ is the time length of a clip and $C^m$ is the channel dimension. $W$ and $H$ are the width and height of the image respectively, and 3 means the number of RGB channels. FineNet generates and retrieves the best dance fragment at each time step,

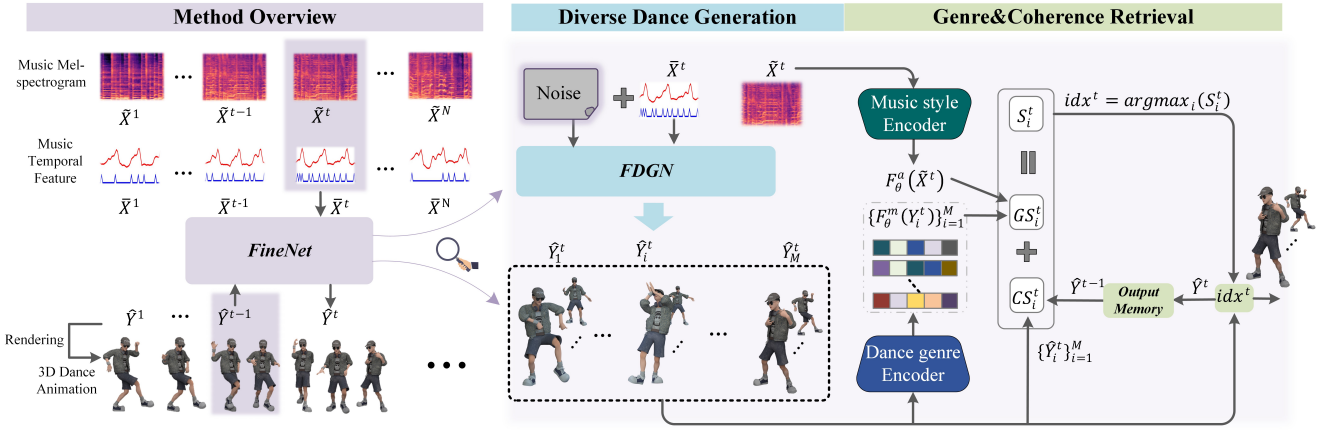


Figure 3. An overview of our framework. FineNet can iteratively generate and choreograph different dance fragments based on the mel-spectrogram and temporal features of music. FineNet consists of a diffusion-based Full body Dance Generation Network (FDGN) and a Genre&Coherence retrieval module. The former is utilized to generate expressive and diverse full-body dance fragments; the latter is designed to retrieve the best matching dance fragments from the generated multiple dance fragments and synthesis them smoothly.

and this process can be formulated as:

$$\begin{aligned} \hat{Y}^1 &= FineNet(\bar{X}^1, \widetilde{X}^1) \\ \hat{Y}^t &= FineNet(\bar{X}^t, \widetilde{X}^t, \hat{Y}^{t-1}), \end{aligned} \tag{1}$$

where $t \in \{2, 3, \cdots, N\}$ is the current time step. $\hat{Y}^t \in \mathbb{R}^{V\times(T\times 3)}$ is the dance action fragment obtained by FineNet at time step t, and $V$ is the number of body and hand joints, "3" represents 3-dimensional axis angle and position of joints.

### 4.2. Diverse Dance Generation

The previous generative models such as VAE[33] and GAN[30], mostly directly link music embedding to dance embedding method, which results in the limited diversity of the generated dances. This is because high-level condensed features of music usually contain insufficient details to guide the network to generate different dances [45].

To generate novel and diverse dances, we employ a diffusion-based model, FDGN. To make the network perceive the music rhythm better, we feed the music temporal feature to FDGN $\bar{X}^t$ instead of $X^t$. Given $\bar{X}^t$ and different noise sampled from $\mathcal{N}(0, I)$, FDGN can generate $M$ distinctive dance fragments, represented as $\{\hat{Y}_i^t\}_{i=1}^M$.

### 4.3. FDGN: Full-body Dance Generation Network

We propose FDGN, a network for generating full body dance motions. Compared to previous models, FDGN can produce more realistic, natural, and expressive hand motions that coordinate well with body motion and music styles. As shown in Figure 4, we use two expert networks to generate body and hand motion separately. This is based on two observations: (1) The range of motion of the body and hand is obviously different and belongs to different feature spaces. Therefore, using a single network to generate

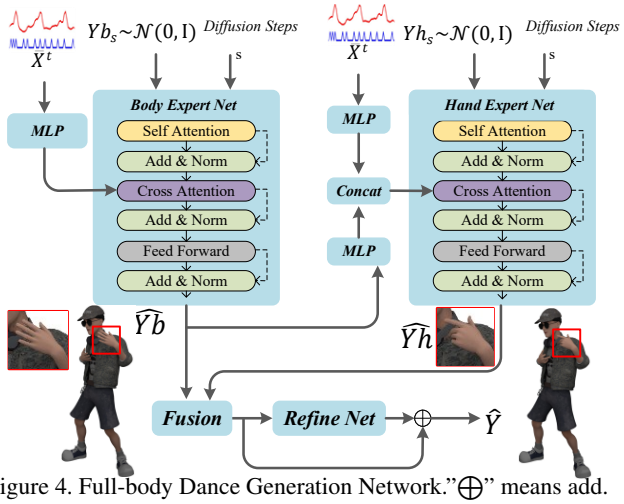


Figure 4. Full-body Dance Generation Network."$\bigoplus$" means add.

full-body dances can result in unnatural hand motions. (2) In dance generation, music rhythm mainly coordinates with the limbs, and hand motion should be consistent with the body and music style. The structure of body/hand expert net is modified by MDM[37] and EDGE[39] . The training process of body expert net can be formulated as:

$$\mathcal{L}_d = \mathrm{E}_{s\in[1,S], Yb_0\sim q(Yb_0)}[\left\|Yb_0 - BEN\left(Yb_s, s, \bar{X}\right)\right\|_2^2], \tag{2}$$

where $Yb_0$ is the label dance data, $BEN(\ )$ means Body Expert Net, $s$ is the number of diffusion steps. The refine net is consisted of a spatial convolution network and a refine gate unit, which is used to assemble two parts naturally.

### 4.4. Genre&Coherent aware Retrieval Module

Thanks to FDGN's diverse generative capabilities, we can generate $M$ candidate dance fragments $\{\hat{Y}_i^t\}_{i=1}^M$ for the $t-th$ music clip $\{\bar{X}^t\}$. To search the genre-matched candidate dance fragment and synthesis them coherently, we propose Genre&Coherent aware Retrieval Module (GCRM),

which calculates the Genre Matching Score (GS) of the current music clip and the candidate dance fragments, and the coherent score of the previous selected dance fragment and the current candidate dance fragments. Finally, we choreograph the complete dance by combining the two scores.

**Genre matching score.** We use GS to evaluate the matching degree of the style of music clip $\widetilde{X^t}$ and the genres of generated candidate dance fragments $\{\hat{Y}_i^t\}_{i=1}^M$. However, music and dance are different data modalities, and it is difficult to calculate the genre similarity directly. To cover this problem, we propose a cross-modal retrieval network.

We utilize two models (one music style encoder $\mathcal{F}_\theta^m(\cdot)$ and one dance genre encoder $\mathcal{F}_\theta^d(\cdot)$) to encode different modalities into one embedding space. The music style encoder is the backbone of AST [9] and the dance genre encoder is the backbone of AGCN [31]. We extract the mel-spectrogram features $\widetilde{X^t}$ of music $X^t$ and send it to the music encoder. For each $\{\widetilde{X^t}\}_{t=1}^N$, we compute its genre-matching score with candidate dance fragments generated by FDGN. We utilize the cosine similarity as the genre matching score. So the genre matching score between $\widetilde{X^t}$ and $\{\hat{Y}_i^t\}_{i=1}^M$ can be formulated as:

$$GS_i^t = s(\widetilde{X^t}, \hat{Y}_i^t) = \frac{\mathcal{F}_\theta^m(\widetilde{X^t})^T \mathcal{F}_\theta^d(Y_i^t)}{||\mathcal{F}_\theta^m(\widetilde{X^t})|| \times ||\mathcal{F}_\theta^d(Y_i^t)||}, \quad (3)$$

We train the above two models by a cosine loss $\mathcal{L}_{cos}$ as a retrieval task, which can be formulated as:

$$\mathcal{L}_{cos} = y(1 - s(\widetilde{X_i}, Y_j)) + (1 - y)(max(0, s(\widetilde{X_i}, Y_j)), \quad (4)$$

where $\widetilde{X_i}$ is the MFCCs feature of music clip, $Y_j$ is a dance fragment. $y = 1$ when $\widetilde{X_i}$ and $Y_j$ are genre-matched, otherwise $y = 0$.

**Coherent score.** We take the L2 distance of the start and end states of the two dance segments as the coherence score. To make the transition smoother, we cut 5 frames of the last temporal clip $(t-1)$ best matching fragment $\hat{Y_b}^{t-1}$ on the tail and cut the start 5 frames of the $M$ candidate dance fragments $\{\hat{Y_i^t}\}_{i=1}^M$ at the current time step $t$. These cut frames are finally filled with a linear interpolation algorithm.

$$\begin{aligned} CS_i^1 &= 0 \\ CS_i^t &= -||Y_b^{t-1}[-5,:] - Y_i^t[5,:]||_2, \end{aligned} \quad (5)$$

where $t \in \{2, 3, \cdots, N\}$ is the time step.

Combining GS and CS, GCRM can find the best $\hat{Y}^t$ from the candidate dance segments $\{Y_i^t\}_{i=1}^M$ as the output at time step $t$:

$$\begin{aligned} idx^t &= \text{argmax}_i\big(\alpha GS_i^t + \beta CS_i^t\big), \\ \hat{Y}^t &:= \hat{Y}_{idx^t}^t, \hat{Y}_{idx^t}^t \in \{Y_i^t\}_{i=1}^M, \end{aligned} \quad (6)$$

where $\alpha$ and $\beta$ are weight parameters. For complete music $X$, FineNet outputs the dance fragments step by step, and the final result is: $\hat{Y} = [\hat{Y}^1, \hat{Y}^2, ..., \hat{Y}^T]$. Furthermore, by choosing different $\hat{Y}^1$ at step 1, FineNet can generate multiple dances with excellent diversity.


Figure 5. Qualitative result comparisons for a Jazz song.

## 5. Experiments

### 5.1. Experimental Setup

**Data Preprocessing.** We spilt FineDance dataset into train, val and test sets in two ways: FineDance@Genre and FineDance@Dancer. The test set of FineDance@Genre includes a broader range of dance genres, but the same dancer appear in train/val/test set. FineDance@Dancer means the train/val/test set divided by different dancers, which test set contains fewer dance genres, yet the same dancer won't appear in different sets. We only reported the results of test set on FineDance@Genre in this paper, the details of dataset split can be found in our supplementary materials. Each music and paired dance are only present in one set. For all the dance fragments, we combine the 3-dim axis angle vector representation for all 52 joints, along with a 3-dim global position vector, resulting in 159-dim motion features. For all the music clips, we use Librosa [27] to extract the 35-dim music temporal features. We also extract the mel-spectrogram of the music clips with Librosa and resized it to $224 \times 224 \times 3$. During extracting audio features, the sampling rate is 76,800Hz and hop size is 512.

**Implementation Details.** In FDGN, we build 3 MLP layers to encode the music and body features. We use the transformer layer as the backbone of the body/hand expert net. The refine net consists of a 1-D convolution layer and a learnable weight parameter. The total epoch, learning rate, and batch size are set as 200, $2e^{-4}$, 2048. In the GCRM, the $\alpha$, $\beta$ are set as 1.0 and 0.5 respectively.

**Evaluation Metrics.** (1) **FID score**. Fréchet inception distance (FID) is widely used to measure how close the distribution of the generated dances is to that of the ground truth.

| Method | FineDance Dataset | | | | | | AIST++ Dataset | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | FID↓ | $FID_h$↓ | Div↑ | $Div_h$↑ | MM↑ | GS↑ | FID↓ | $FID_h$↓ | Div↑ | $Div_h$↑ | MM↑ | GS↑ |
| Ground Truth | / | / | 6.28 | 3.48 | / | 0.71 | / | / | 9.07 | / | / | 0.77 |
| ChoreoMaster# | 1.92 | 0.61 | **6.18** | 3.33 | 12.20 | 0.39 | 2.21 | / | 9.38 | / | 30.71 | 0.71 |
| DanceRevolution* | 7.44 | 3.21 | 3.92 | 1.80 | 4.22 | 0.30 | 6.05 | / | 7.67 | / | 12.88 | 0.75 |
| DeepDance* | 5.77 | 1.95 | 5.07 | 3.50 | 0.73 | 0.36 | 25.78 | / | 8.98 | / | 3.08 | 0.73 |
| Bailando* | 4.77 | 2.24 | 3.09 | 1.12 | 2.33 | 0.29 | 17.45 | / | 9.44 | / | 2.58 | 0.74 |
| **FineNet (w/o. FDGN)** | 1.90 | 1.20 | 5.87 | 3.40 | 12.86 | 0.62 | 2.11 | / | 9.13 | / | 31.09 | 0.79 |
| **FineNet** | **1.66** | **0.48** | 5.99 | **3.59** | **16.72** | **0.74** | **2.05** | / | **9.94** | / | **33.67** | **0.80** |

Table 2. Compared with SOTAs. # means we reproduce the code, * means we use the published code.

Similar to Lee *et al.*[19], we trained a style classifier in our dataset to extract motion features, and then use the features to calculate FID. (2) **Diversity**. We follow Lee *et al.* [19] to evaluate the average feature distance between generated dances for different input music. The same feature extractor used in FID is used again. (3)**Hand FID score** and **Hand Diversity**. Similarly, we extract hand motion features and calculate the FID and diversity for hand motion. (4) **Multimodality**. We follow Lee *et al.*[19] to evaluate the average feature distance between the 10 choreography versions of every music. This metric measures the model's ability to generate different dances for the same music. (5) **Genre Matching Score.** We evaluate the average genre matching score between generated dance and the input music using the genre matching score calculation network mentioned in section 4.4. The genre matching score is defined as Eq. (3).

## 5.2. Quantitative and Qualitative Evaluation

**Compared methods.** We compare our method with several generation-based methods and one synthesis-based method. ChoreoMaster [5] is a dance synthesis method with style embedding network and graph-based motion synthesis. Since the code is not available, we reproduced the code according to the paper. DanceRevolution [14] is a generation method using Tansformer to generate long-term dances. DeepDance [10] is a generation method with a GAN-based cross-modal association framework. Bailando [33] is a generation method with VAE and an actor-critic generative pre-trained transformer model. All the methods are tested on FineDance and AIST++ to evaluate the comprehensive choreography ability of FineNet.

**Results and analysis.** For all methods, we generate 10 versions of the dance for each song in the test set, and take the paired real dance as the Ground Truth. As shown in Table 2, our method gets the best performance in all evaluation metrics except Diversity on both datasets. ChoreoMaster gets better on Diversity because it must retrieve all the fragments in the whole training set but no new motion is created. Specifically, on Multimodality (MM), FineNet gains 4.52 improvements compared to ChoreoMaster. On GS, FineNet is 0.03 higher than the ground truth. These two metrics show FineNet can generate diverse and genre-matched dance. Besides, FineNet gets the highest Hand FID and Hand Diversity which shows our method can generate real and diverse hand motion.

Qualitative results are shown in Figure 5. DanceRevelution, DeepDance and Bailando generated almost identical stiff motions (refer to row 1 and row 2). ChoreoMaster generates classical dance motions in the last two images that do not match the Jazz music. But the generated dance motions of FineNet are diverse and in line with the Jazz style. The hand motion generated by FineNet W/ FDGN is more real. The corresponding videos refer to supplementary material.

## 5.3. Ablation Study

We replace different components of our method, and conduct experiments to demonstrate the effectiveness of each component on the test set of FineDance.

**Infulence of the FDGN.** In order to verify whether FDGN can make hand movements more natural and flexible. We modified FDGN as a single network that generates body and hand motion simultaneously. We then employed $FID_h$ and $DIV_h$ metrics to evaluate the realism and diversity of the resulting hand motion. As shown in Table2, all the quantitative metrics for FineNet w/o. FDGN becomes worse on both datasets. Specifically, on Hand Fid and Hand Diversity, FineNet w/o. FDGN drops by 0.72 and 0.14. The huge increase in $FID_h$ is mainly because FDGN generates hand movements with complete body information, making the hand motions more coordinated with the body.

| Strategy | FID ↓ | Diversity ↑ | MM ↑ | GS ↑ |
|---|---|---|---|---|
| Ground Truth | / | 6.28 | / | 0.71 |
| FDGN-G | 2.85 | 6.22 | 12.02 | 0.27 |
| FDGN-C | 1.68 | **6.34** | 11.51 | 0.46 |
| **FineNet** | **1.66** | 5.99 | **16.72** | **0.74** |

Table 3. Ablation study on different choreography strategies.

**Choreography Strategy.** To verify the proposed generative-synthesis strategy, we use the following strategies to generate 3D dances: directly generating long-term 3D dances by FDGN (FDGN-G); generating 4s dance fragments according to the music clip by FDGN and concatenating these fragments together (FDGN-C), as Table 3 shows.

FineNet performs better than others in FID, MM, and GS. For FDGN-G, a long music clip will cause the FID and GS to drop dramatically. FineNet achieves a lower Diversity score than FDGN-C because while retrieving the GCRM gives up some dances whose genre doesn't match. For MM and GS, FineNet increases by 5.21 and 0.28 compared to FDGN-C, which demonstrates that GCRM can increase the ability to generate diverse and genre-matched dance for the same song.

| Generation Model | FID ↓ | Diversity ↑ | MM ↑ | GS ↑ |
|---|---|---|---|---|
| DeepDance | 4.91 | 4.47 | 1.15 | 0.32 |
| DanceRevolution | 6.99 | 3.51 | 6.64 | 0.37 |
| **FDGN** | **1.66** | **5.99** | **16.72** | **0.74** |

Table 4. Ablation study on different generation networks.

**Influence of the generation model.** We choose DanceRevolution, DeepDance as the backbone of the generative model, and get the whole 3D dances through our GCRM. All the results are shown in Table 4. Our FDGN performs best, which shows excellent dance generation ability. Besides, compared the results of Table 2 and Table 4, where DeepDance and DanceRevolution are inserted into our method in Table 4, and their performances can be improved on MM, FID and GS. These results show our GRCM is scalable.

### 5.4. User Study

We invite 30 participants, including 15 dancers. Everyone watches 5 dances with a duration of 32 seconds. They are asked to evaluate the dances in 2 aspects: Effective duration and Artistry.

**Effective duration.** Each participant has to judge the time interval from the beginning until abnormal actions occur, such as abnormal shaking, motion freezing, or excessive joint distortion. As Table 5 shows, synthesis methods generally performs better than generation methods, generation methods can only generate effective dances for less than 10 seconds. These results show that generation methods are limited in effective duration. But, ours is a generative-synthesis method, and achieves the best performance.

**Artistry.** We measure the artistry of the dances from the following dimensions: *genre matching (GM), rhythm matching (RM), diversity (Div.), and comprehensive artistry (CA).* The full score of every dimension is 100. As shown in Table 6, synthesis methods generally perform better than generation methods, and our method gets a 77.0 average score which clearly surpasses the baselines. However, compared to the ground truth, all the methods still need to improve especially in rhythm matching and Comprehensive artistry.

| Method | Duration ↑ | Type |
|---|---|---|
| Ground Truth | 30.7s | / |
| ChoreoMaster | 26.9s | synthesis |
| DanceRevolution | 5.4s | generation |
| DeepDance | 2.2s | generation |
| Baildando | 3.1s | generation |
| **FineNet (Ours)** | **28.2s** | generation&synthesis |

Table 5. User study on effective duration.

| Method | GM | RM | Div. | CA | Avg |
|---|---|---|---|---|---|
| Ground Truth | 94.1 | 96.3 | 95.5 | 97.9 | 96.0 |
| ChoreoMaster | 64.2 | 59.7 | 77.2 | 69.8 | 67.7 |
| DanceRevolution | 33.2 | 30.7 | 47.4 | 31.4 | 36.7 |
| DeepDance | 17.2 | 22.1 | 24.4 | 25.6 | 22.3 |
| Bailando | 23.2 | 33.5 | 39.4 | 21.6 | 30.1 |
| **FineNet (Ours)** | **79.9** | **70.3** | **81.2** | **76.4** | **77.0** |

Table 6. Result of the user study. Avg denotes the average score.

## 6. Conclusions

In this paper, we propose a large-scale, high-quality 3D dance dataset (FineDance) for music-driven dance generation, which records professional and abundant dance genres with accurate and fine-grained hand motions. Meanwhile, we also propose a choreography Network (FineNet), which can generate multiple diverse genre-matched dances with flexible hand movements. Furthermore, we propose a new metric to evaluate the genre matching degree between music and dance. Quantitative and qualitative results show that FineNet can generate long-term, diverse, and genre-matched dances from given music.

## Acknowledgment

We would like to express our sincere gratitude to Yan Zhang (https://yz-cnsdqz.github.io/) and Yulun Zhang (https://yulunzhang.com/) for their invaluable guidance and insights during the course of our research.

This work was supported in part by the Shenzhen Key Laboratory of next generation interactive media innovative technology (No.ZDSYS20210623092001004), in part by the China Postdoctoral Science Foundation (No.2023M731957), in part by the National Natural Science Foundation of China under Grant 62206153, Young Elite Scientists Sponsorship Program by CAST (No. 2023QNRC002)

## References


[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3d human motion modelling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7144–7153, 2019.

[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.

[3] Okan Arikan and David A Forsyth. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)*, 21(3):483–490, 2002.

[4] Alexander Berman and Valencia James. Kinetic imaginations: exploring the possibilities of combining ai and dance. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, page 2431–2437, 2015.

[5] 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.

[6] Joao P Ferreira, Thiago M Coutinho, Thiago L Gomes, José F Neto, Rafael Azevedo, Renato Martins, and Erickson R Nascimento. Learning to dance: A graph convolutional adversarial network to generate realistic dance motions from audio. *Computers & Graphics*, 94:11–21, 2021.

[7] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*, pages 4346–4354, 2015.

[8] Partha Ghosh, Jie Song, Emre Aksan, and Otmar Hilliges. Learning human motion models for long-term predictions. In *2017 International Conference on 3D Vision (3DV)*, pages 458–466. IEEE, 2017.

[9] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[11] Judith Lynne Hanna. *To dance is human: A theory of nonverbal communication*. University of Chicago Press, 1987.

[12] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, 2023.

[13] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *arXiv preprint arXiv:2305.11003*, 2023.

[14] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. In *International Conference on Learning Representations*, 2021.

[15] Jae Woo Kim, Hesham Fouad, and James K Hahn. Making them dance. In *AAAI Fall Symposium: Aurally Informed Performance*, volume 2, 2006.

[16] Tae-hoon Kim, Sang Il Park, and Sung Yong Shin. Rhythmic-motion synthesis based on motion-beat analysis. *ACM Transactions on Graphics (TOG)*, 22(3):392–401, 2003.

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[18] Alexis Lamouret and Michiel van de Panne. Motion synthesis by example. In *Computer Animation and Simulation'96*, pages 199–212. Springer, 1996.

[19] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[20] Minho Lee, Kyogu Lee, and Jaeheung Park. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 62(3):895–912, 2013.

[21] Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:1272–1279, 06 2022.

[22] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13381–13392, 2021.

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34:248:1–248:16, 2015.

[24] Kovar Lucas, Gleicher Michael, and Pighin Frédéric. Motion graphs. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, pages 473–482, 2002.

[25] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. *arXiv preprint arXiv:2304.01186*, 2023.

[26] Adriano Manfrè, Ignazio Infantino, Filippo Vella, and Salvatore Gaglio. An automatic system for humanoid dance creation. *Biologically Inspired Cognitive Architectures*, 15:1–9, 2016.

[27] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

[28] Ferda Ofli, Engin Erzin, Yücel Yemez, and A Murat Tekalp. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia*, 14(3):747–759, 2011.

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and

Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[30] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020.

[31] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019.

[32] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Dancing-to-music character animation. In *Computer Graphics Forum*, volume 25, pages 449–458. Wiley Online Library, 2006.

[33] 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.

[34] 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, 2021.

[35] Taoran Tang, Jia Jia, and Mao Hanyang. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. pages 1598–1606, 10 2018.

[36] Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, and Xiu Li. Flag3d: A 3d fitness activity dataset with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22106–22117, 2023.

[37] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Amit H Bermano, and Daniel Cohen-Or. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

[38] Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. Edge: Editable dance generation from music. *arXiv preprint arXiv:2211.10658*, 2022.

[39] 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.

[40] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30:5998–6008, 2017.

[42] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8639–8648, 2018.

[43] Yanzhe Yang, Jimei Yang, and Jessica Hodgins. Statistics-based motion synthesis for social conversations. In *Computer Graphics Forum*, volume 39, pages 201–212. Wiley Online Library, 2020.

[44] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 744–752, 2020.

[45] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022.

[46] Yan Zhang, Michael J Black, and Siyu Tang. We are more than our joints: Predicting how 3d bodies move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3372–3382, 2021.

[47] Yan Zhang and Siyu Tang. The wanderings of odysseus in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20481–20491, 2022.

[48] Yachao Zhang, Yuan Xie, Cuihua Li, Zongze Wu, and Yanyun Qu. Learning all-in collaborative multiview binary representation for clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[49] 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.