GCDance: Genre-Controlled Music-Driven 3D Full Body Dance Generation

Xinran Liu, Xu Dong, Shenbin Qian, Diptesh Kanojia,
Wenwu Wang, *Fellow, IEEE,* and Zhenhua Feng*, *Senior Member, IEEE*

***Abstract*—Music-driven dance generation is a challenging task as it requires strict adherence to genre-specific choreography while ensuring physically realistic and precisely synchronized dance sequences with the musical beats and rhythm. Although significant progress has been made in music-conditioned dance generation, most existing methods struggle to convey specific stylistic attributes in generated dance. To bridge this gap, we introduce a diffusion-based framework for genre-specific 3D full-body dance generation, conditioned on both music and descriptive text. To effectively incorporate genre information, we develop a text-based control mechanism that maps input prompts, either explicit genre labels or free-form descriptive text, into genre-specific control signals, enabling precise and controllable text-guided generation of genre-consistent dance motions. Furthermore, to strengthen the alignment between music and textual conditions, we leverage the features of a music foundation model, facilitating coherent and semantically aligned dance synthesis. Last, to balance the objectives of extracting text-genre information and maintaining high-quality generation results, we propose a novel multi-task optimization strategy. This effectively balances competing factors such as physical realism, spatial accuracy, and text classification, significantly improving the overall quality of the generated sequences. Extensive experiments on the FineDance, AIST++, and PopDanceSet datasets demonstrate the superiority of GCDance over the existing state-of-the-art approaches. The source code and demonstration videos are available at https://xinranliu7715.github.io/gcdance/.**

***Index Terms*—3D human dance, music to dance generation, diffusion model, controllable generation, multi-task learning.**

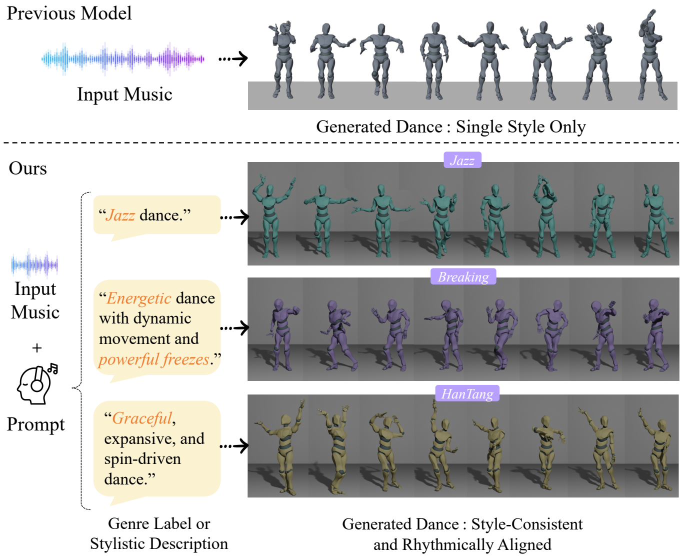
Previous Model
Input Music
Generated Dance : Single Style Only
Ours
Jazz
"*Jazz* dance."
Breaking
"*Energetic* dance with dynamic movement and *powerful freezes*."
HanTang
"*Graceful*, expansive, and spin-driven dance."
Input Music
+
Prompt
Genre Label or Stylistic Description
Generated Dance : Style-Consistent and Rhythmically Aligned

Fig. 1: Conditioned on music and a genre-descriptive prompt, GCDance generates 3D dance motions that are temporally aligned with the rhythm and semantically consistent with the textual instruction.

I. Introduction

DANCING is a universal form of cultural expression and a powerful medium for conveying emotions. However, choreography is an artistic skill that demands years of training. During the choreographic process, the body movements of the choreographer need to be aligned with the musical rhythm while reflecting the stylistic characteristics of a specific dance genre [1]. As a result, the use of AI for music-driven choreography shows promising research potential.

In recent years, numerous deep-learning-based methods have been proposed for the task of dance generation. Early dance generation approaches often rely on autoregressive models that directly predict future dance movements from the past motion sequences [2], [3], but they frequently encounter challenges such as motion freezing during long-term generation. To mitigate this issue, Vector-Quantized Variational AutoEncoder (VQ-VAE) based methods [4], [5] introduce a discrete codebook of motion units, effectively stabilizing long-range motion. Nevertheless, the reliance on a fixed latent vocabulary inherently restricts the diversity and expressiveness of generated dances [6]. More recently, diffusion models [7] have shown remarkable performance in various generation tasks. Unlike methods that rely on predefined seeds or fixed latent vocabularies, diffusion models iteratively refine noise into coherent outputs, thereby capturing a broader space of potential motions. These approaches [8], [9], [10] greatly enhance both diversity and expressiveness of dance motions generated. However, existing approaches often struggle to convey specific stylistic attributes. Although these methods can generate a single style of dance for a given piece of music, they may lead to mismatches between the generated motion and the musical style, or may fail to produce dances that align with a user-intended genre.

To address these limitations, we propose GCDance, a genre-

X. Liu and D. Knojia are with the School of Computer Science and Electronic Engineering, University of Surrey, Guildford GU2 7XH, UK (e-mail: xinran.liu@surrey.ac.uk; d.kanojia@surrey.ac.uk)

X. Dong is with the Department of Music and Media, University of Surrey, Guildford GU2 7XH, UK (e-mail: x.dong@surrey.ac.uk)

S. Qian is with the Department of Informatics, University of Oslo, 0316 Oslo, Norway (email: shenbinq@ifi.uio.no)

W. Wang is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, UK (e-mail: w.wang@surrey.ac.uk).

Z. Feng is with the School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China (e-mail: fengzhenhua@jiangnan.edu.cn)

*Corresponding Author

controllable 3D full-body dance generation model conditioned on both music and text. GCDance focuses on generalization to high-fidelity motions while maintaining controllability. Specifically, we introduce a classification-based control mechanism utilizing explicit genre labels or descriptive natural language prompts as input. The textual input is first classified to determine its corresponding dance genre, and subsequently encoded into control signals to guide the generation process, enabling the model to modulate the generated dance style accordingly. With the introduction of the text as an additional conditioning modality, aligning it with the music representation is critical for achieving consistent and controllable dance generation. However, the majority of current approaches are based solely on hand-crafted music features [11], [12], which are typically low-level and inadequate for modeling the complex and nuanced correlations between music and textual descriptions. To achieve a better alignment between these multimodal signals, we integrate hand-crafted features with deep representations from the Wav2CLIP music foundation model [13]. By mapping audio and text into a shared embedding space, Wav2CLIP enables cross-modal alignment that captures musical semantics and genre attributes with a single representation for motion control, thereby enabling the model to generate stylized dance motions conditioned on various textual prompts.

Apart from the above genre-controlled mechanism, achieving robust dance generation inherently involves multiple objectives. This requires the model to balance goals such as spatial accuracy, temporal coherence, and genre control. In practice, these objectives may conflict with each other, leading to trade-offs in the generated motions. For example, increasing motion diversity can reduce fidelity and coherence, and highly realistic sequences may still fail to reflect the intended genre style. Existing approaches typically consolidate these competing objectives into a single loss function with manually tuned weights [8], [14], often leading to suboptimal trade-offs among different aspects of motion quality. To achieve a dynamic balance among these tasks and enhance the performance of generated dances, we adopt a Multi-Task Learning (MTL) framework that jointly optimizes multiple objectives such as motion quality, velocity constraints, foot contact consistency, and genre classification. By assigning a distinct objective function to each requirement, our method dynamically adjusts the training process and ultimately improves motion quality, achieving state-of-the-art performance across multiple quantitative evaluation metrics. In addition, our model is trained on a dataset with 52-joint full-body representations, which include detailed hand movements. This richer skeletal representation further enhances the realism and expressiveness of the generated dances by capturing fine-grained motion details that are neglected in previous studies.

In summary, our main contributions include:

- We introduce a diffusion-based multi-genre dance generation model, namely GCDance. It enables controllable dance generation by conditioning on both music and textual prompts.
- To enhance cross-modal alignment, GCDance leverages a pretrained music foundation model that captures both high-level semantic cues and low-level audio details for more coherent and expressive dance generation.
- We introduce a novel multi-task learning framework that jointly optimizes diverse objectives for a more balanced model training.
- Extensive experiments on the FineDance, AIST++, and PopDanceSet datasets show that GCDance consistently outperforms existing approaches on various metrics.

II. Related work

A. Music Driven Dance Generation

Early studies [15], [16] treat this task as a similarity-based retrieval problem, where motion segments are selected from a predefined database based on the input music. These approaches inherently limit the diversity and creativity of the generated dances. To mitigate this, deep learning methods reframe the task as motion prediction using architectures such as Convolutional Neural Network (CNN) [17], Recurrent Neural Network (RNN) [18], [19], and Transformers [20], [21], [22]. However, these frame-by-frame prediction methods are often prone to error accumulation and motion freezing [23].

Recent research has shifted to the generative pipeline. TM2D [5] quantizes 3D motion with a VQ-VAE and uses a cross-modal Transformer conditioned on music and text to predict motion tokens and generate dance segments. Bailando [4] learns a discrete motion vocabulary and frames dance synthesis as music-conditioned token prediction, using a reinforcement learning evaluator that derives beat-alignment rewards from the audio to enhance rhythmic synchrony. These systems make progress, yet they are complex and often rely on handcrafted musical features such as Mel-Frequency Cepstral Coefficient (MFCC), chroma features, and one-hot beat indicators, which miss the intricate detail needed for precise music–movement correlation. EDGE [8] proposes a single transformer-based diffusion model for generating dance driven by music, paired with a Jukebox feature extractor and strong editing capabilities. POPDG [24] builds on iDDPM [25] with Jukebox features and introduces a space augmentation algorithm to achieve a balance between generation quality and diversity. Nevertheless, these approaches remain insufficient to achieve genre controllability. The potential of textual semantics is still underexplored, resulting in unstable style control and limited genre diversity.

B. Diffusion Models

Diffusion models [7], [25] have advanced rapidly in recent years [26] and are widely adopted in multiple domains, including image generation [27], [28], audio synthesis [29], [30], [31], and text generation [32], [33]. For conditional generation, existing approaches often employ classifier guidance [34], [35] or classifier-free guidance [36], [37] to improve sample quality, both of which operate at inference time and apply to pretrained diffusion models without fine-tuning.

Moreover, the strong capacity of diffusion models for controllable generation is gaining significant attention. Blended Diffusion [38] introduces a model for text-conditional image generation. It leverages CLIP [39] to guide the diffusion process, ensuring the resulting images align with the target

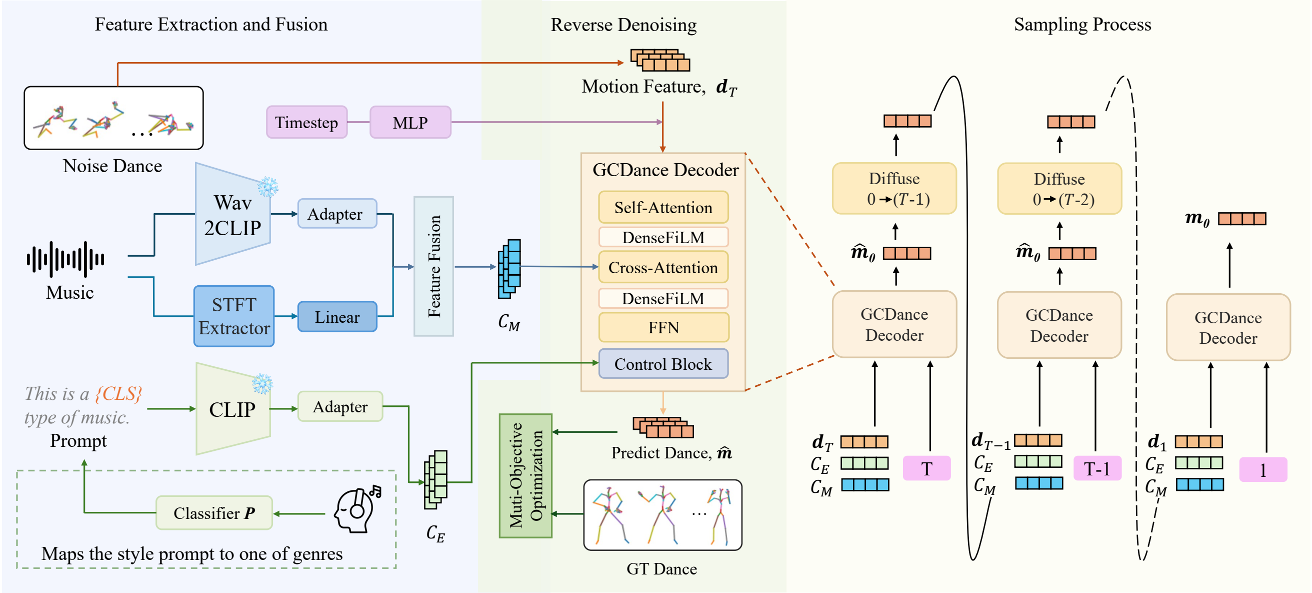

Fig. 2: An overview of GCDance. Left: the multimodal inputs and feature extraction. Middle: the training process at a given diffusion timestep t. Right: the sampling process, which iteratively generates a sequence of dance motions.

prompt. GMD [40] adapts diffusion for trajectory generation, incorporating spatial constraints to enhance the correspondence between spatial information and local poses. Alexanderson *et al.* [41] introduce an audio-driven method for gesture and dance generation, featuring controls for style and expressive intensity. However, this method is limited to only four genres. Note that dance motion generation is inherently complex and further challenged by data scarcity [8]. In our work, we propose a diffusion-based method that generates 16 different dance genres from music and allows for text-based control over various dance types.

C. Multi-Task Learning

Multi-Task learning (MTL) trains related tasks simultaneously using a shared representation. Although early MTL methods sometimes underperform single-task models [42], recent approaches have overcome these issues. For example, MTAN [43] is a multi-task learning architecture that uses dynamic weight averaging with task-specific feature-level attention by employing a shared network and soft-attention modules without preset weighting schemes. Similarly, an impartial MTL was proposed in [44], which uses distinct strategies for shared and task-specific parameters. In addition, Nash-MTL [45] re-frames the gradient combination as a bargaining game, using the Nash Bargaining Solution [46] to negotiate a joint update direction among tasks. To improve training stability, Aligned MTL was developed [47], which aligns the orthogonal components of gradient systems according to their condition number. Furthermore, a Bayesian gradient aggregation method was introduced to model uncertainty over task-specific parameters and gradients [48]. These advances have been widely applied in various fields in computer vision [49], [50] and natural language processing [51], [52].

Combining different training objectives is common in dance generation. However, existing approaches typically consolidate these competing objectives into a single loss function with manually tuned weights [8], [14], rather than weights learned by parametric heuristics. This often leads to suboptimal trade-offs among different aspects of motion quality. To address this issue, we propose a novel multi-objective training strategy that integrates parametric loss heuristics like Nash MTL and Aligned MTL to optimize these training objectives.

III. The Proposed GCDance Method

In this section, we present the details of the proposed GCDance method, introducing its overall architecture and key components. The diffusion preliminaries of our approach are provided in the supplementary material.

A. The GCDance Architecture

The overall architecture of GCDance is illustrated in Fig. 2. We define three modalities in the framework: dance motion, music, and textual prompt. Each modality is turned into an informative representation as detailed below.

Given a long music-dance pair, we first divide it into N 4-second segments. For each segment, we uniformly sample k frames from the corresponding dance motion and music clip.

To represent **dance motion**, we employ the Skinned Multi-Person Linear (SMPL) format [53], and define three primary components: (1) Human joint positions: The 52 joint positions are transformed into a 312 (i.e. 52 × 6) dimensional rotation representation with 6 degrees of freedom (DOF), denoted as $\boldsymbol{p} \in \mathbb{R}^{312}$. (2) Root translation: A 3D vector describes the global translation of the root joint. (3) Foot-ground contact: Following EDGE [8], a 4D contact label $\boldsymbol{f} \in \mathbb{R}^4$ encodes the binary heel and toe contact states for each foot. In total, the complete pose sequence is represented as $\boldsymbol{m} \in \mathbb{R}^{k\times 319}$, where k represents the number of frames.

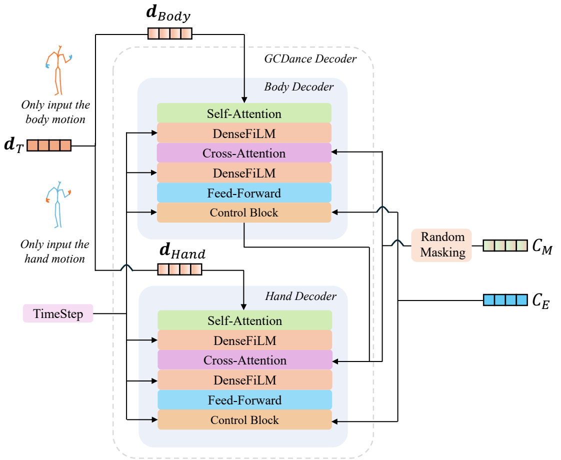

Fig. 3: The decoder of GCDance.

For **music representations**, existing approaches typically rely on hand-crafted musical features, overlooking recent advances in music foundation models, which have shown strong potential for capturing nuanced representations of music. To address this limitation, GCDance integrates music embeddings extracted from a pretrained music foundation model with hand-crafted music features, effectively utilizing both high-level semantic insights and low-level temporal specifics to improve the fidelity of the generated dance sequences. For high-level representations, we adopt Wav2CLIP [13] as the music encoder. Wav2CLIP is an audio-visual correspondence model that distills from the CLIP framework [39]. It is trained to predict CLIP-style embeddings from raw audio by aligning them with frozen vision-based representations extracted from videos. For hand-crafted music features, we employ Short-time Fourier Transform (STFT) that captures fine-grained temporal-frequency features in music signals [54]. In GCDance, we extract STFT features using the Librosa toolbox [55].

For **text representations**, our goal is to establish a free form text guided dance generation framework. However, the absence of text–dance paired data in existing datasets presents a significant limitation. To overcome this issue, we construct a dance genre description dataset and develop a genre classifier $\boldsymbol{P}$ that maps free-form textual descriptions C_{desc} to genre $\hat{g}$. Comprehensive details of this dataset are provided in the supplementary materials.

To evaluate the genre classifier, we compute the binary cross-entropy (BCE) loss which measures the divergence between the predicted distribution $\hat{g}$ and the ground-truth genre label g for each music–text pair:

$$\hat{g} = P(C_{\text{desc}}), \tag{1}$$

$$L_C = BCE\left(\hat{g}, g\right) \tag{2}$$

Based on the predicted genre $\hat{g}$, we apply a prompt learning strategy [56] to transform the discrete label into a complete textual prompt, thereby providing genre-related semantic information to guide the generation process. For example, given the genre label “Jazz,” the generated sentence is “This is a Jazz type of music.”. CLIP is then employed to encode this prompt into a semantic embedding, denoted as C_E, which captures genre-specific textual semantics aligned with the user’s input.

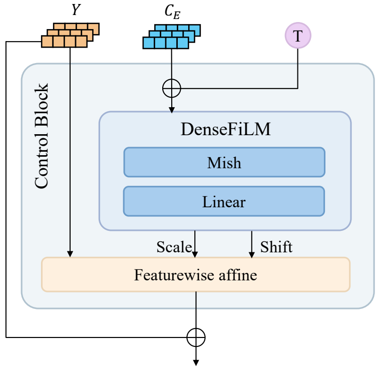

Fig. 4: The control module of GCDance.

Last, GCDance takes the inputs as noise slice $\boldsymbol{d}_T$, music condition C_M, text genre embedding C_E, and diffusion timestep t. These inputs are then processed by a Transformer-based denoising model. As depicted in Fig. 3, we use two specialized expert downsampling modules inspired by [14], each tasked with independently modeling the distributions of body and hand motion. The approach leverages the distinct kinematics and degrees of freedom between body and hands. By employing two specialized branches that learn component-specific representations, it improves local kinematic fidelity and overall expressiveness in the generated motion.

More specifically, the motion sequences are processed in parallel by two Transformer-based networks. Each network includes stacked self-attention and cross-attention blocks, multi-layer perceptrons, and Feature-wise Linear Modulation (FiLM) layers [57]. The output features from the body decoder are injected into the cross-attention layer of the hand decoder to help capture the relationship between body and hand movements effectively. Despite this design, a significant domain gap remains between the conditioning feature space and the motion feature space. To mitigate this gap, we introduce a lightweight adapter that processes the extracted music and text representations and maps them into a motion-aligned latent space. Additionally, to integrate the music conditioning, the decoder applies cross-attention in which queries attend to music keys and values derived from the projected music embeddings, following [58].

Genre-Controllability. As shown in Fig. 4, the control module integrates genre information into the generation process at each diffusion timestep through a FiLM layer. FiLM modulates the intermediate activations of the network through affine transformations conditioned on external inputs, enabling dynamic adaptation of the contextual signal representations.

In GCDance, we use the output from the previous network layer, denoted as Y, along with a genre embedding C_E as inputs to the control module. The genre embedding is conditioned on the current diffusion timestep and then used to derive the FiLM modulation parameters as follows:

$$\gamma = \theta_w(\alpha(C_E)), \quad \varepsilon = \theta_b(\alpha(C_E)) \tag{3}$$

$$FiLM_t(Y) = \gamma \odot Y + \varepsilon, \tag{4}$$

where α denotes a text-embedding adapter that refines the representation, $\odot$ denotes element-wise multiplication, and θ_w

and θ_b are learned linear projections.

B. Multi-Objective Training

Training Objective. The training process involves five objectives. We employed the loss function $\mathcal{L}_S$ from DDPM as our primary objective, defined as:

$$\mathcal{L}_{\text{S}} = \mathbb{E}_{\boldsymbol{m}_0,t}\left[\|\boldsymbol{m}_0 - f_{rev}(\boldsymbol{d}_t,t,C_M,C_E)\|_2^2\right] \tag{5}$$

In addition, to generate fluent and physically-plausible motion sequences, we additionally integrate several auxiliary losses common to motion generation tasks, such as EDGE [8] and Motion Diffusion Model (MDM) [6]. These auxiliary losses are designed to align three key aspects: joint positions (Equ. 6), velocities (Equ. 7), and foot contact (Equ. 8). Following prior work [6], the forward kinematic function $FK(\cdot)$ is utilized to derive joint positions from joint angles, which facilitates the joint loss calculation:

$$\mathcal{L}_{\text{J}} = \frac{1}{n}\sum_{j=1}^{n}\left\|FK\left(\boldsymbol{m}^j\right) - FK\left(\hat{\boldsymbol{m}}^j\right)\right\|_2^2 \tag{6}$$

where j denotes the frame index and $\hat{\boldsymbol{m}}^j$ is the predicted pose for the j-th frame. Furthermore, velocity and acceleration are computed, leading to the velocity loss:

$$\mathcal{L}_{\text{V}} = \frac{1}{n-1}\sum_{j=1}^{n-1}\left\|\left(\boldsymbol{m}^{j+1} - \boldsymbol{m}^j\right) - \left(\hat{\boldsymbol{m}}^{j+1} - \hat{\boldsymbol{m}}^j\right)\right\|_2^2 \tag{7}$$

Finally, we utilize the contact loss $\mathcal{L}_{\text{F}}$, leveraging binary foot-ground contact labels to optimize foot contact consistency during motion generation:

$$\mathcal{L}_{\text{F}} = \frac{1}{n-1}\sum_{j=1}^{n-1}\left\|\left(FK\left(\hat{\boldsymbol{m}}^{j+1}\right) - FK\left(\hat{\boldsymbol{m}}^j\right)\right)\cdot\hat{\boldsymbol{b}}^j\right\|_2^2 \tag{8}$$

where $\hat{\boldsymbol{b}}^j$ is the predicted binary foot-ground contact label.

To balance multiple training objectives and address the optimization challenges such as conflicting or dominating gradients, we propose a multi-objective training strategy below:

$$\mathcal{L} = \tau\left(\mathcal{L}_{\text{S}}, \mathcal{L}_{\text{J}}, \mathcal{L}_{\text{V}}, \mathcal{L}_{\text{F}}, \mathcal{L}_{\text{C}}\right). \tag{9}$$

This strategy relies on a heuristic function τ that combines five distinct losses into a single optimization objective. The goal is to find a parameter set θ that minimizes the overall aggregation loss:

$$\Delta\theta = \min_{\theta}\sum_{i=1}^{T}\mathcal{L}_i\left(\theta_i\right) \tag{10}$$

where T denotes the number of loss components, and $\mathcal{L}_i(\theta_i)$ represents the i-th loss function.

MTL Training Strategy. In our implementation, we explore two different heuristics, including Nash MTL [45] and Aligned MTL [47], to learn the parameter set θ.

Nash MTL is designed to compute an update vector $\Delta\theta$ that integrates the task-specific gradients g_i, while ensuring that $\Delta\theta$ remains within an ϵ-radius ball centered at zero, denoted by B_ϵ. This is formulated as the following optimization problem:

$$\arg\max_{\Delta\theta\in B_\epsilon}\Sigma_i log(\Delta\theta^\intercal g_i) \tag{11}$$

The optimal solution to this problem is (up to scaling) $\Sigma_i\alpha_i g_i$, where $\alpha \in \mathbb{R}_+^K$ is the solution to $G^\intercal G\alpha = 1/\alpha$ with the reciprocal taken element-wise. The complete Nash MTL algorithm is specified as follows:

Algorithm 1 Nash-MTL

```
Require: Initial parameter vector $\theta^{(0)}$, differentiable loss functions $\{l_i\}_{i=1}^K$, learning rate $\eta$
for $t = 1, \ldots, T$ do
    Compute task gradients $g_i^{(t)} \leftarrow \nabla_{\theta^{(t-1)}} l_i$
    Form matrix $G^{(t)}$ with columns $g_i^{(t)}$
    Solve for $\alpha$: $(G^{(t)})^\intercal G^{(t)}\alpha = 1/\alpha$ to obtain $\alpha^{(t)}$
    Update parameters: $\theta^{(t)} \leftarrow \theta^{(t-1)} - \eta G^{(t)}\alpha^{(t)}$
end for
return $\theta^{(T)}$
```

Aligned MTL is a method that aligns the principal components of the gradient matrix to enhance training stability. As formulated in Equ. 12, it minimizes the Frobenius distance between the original gradient matrix G and its aligned version $\hat{G}$. A key condition imposed by this equation is the orthogonality of $\hat{G}$, which signifies that the product of $\hat{G}$ and its transpose is the identity matrix. This orthogonality decorrelates task update directions and normalizes their scales, which improves the conditioning of the gradient system and stabilizes optimization.

$$\min_{\hat{G}}\|G - \hat{G}\|_F^2 \quad s.t. \quad \hat{G}^\intercal\hat{G} = I \tag{12}$$

$$\hat{G} = \sigma U V^\intercal = \sigma G V \Sigma^{-1} V^\intercal \tag{13}$$

Equ. 13 outlines the approach, where $\hat{G}$ is determined through singular value decomposition (SVD). In this procedure, the matrix G is decomposed into three distinct components: U, Σ, and $V^\intercal$. U and V form orthogonal bases, and Σ is a diagonal matrix containing the non-negative singular values of G, typically arranged in descending order. The complete Aligned MTL algorithm is formally detailed below:

Algorithm 2 Aligned-MTL

```
Require: Gradient matrix $G \in \mathbb{R}^{|\theta|\times T}$, task importance $w \in \mathbb{R}^T$
Compute $M \leftarrow G^\intercal G$
Perform eigen-decomposition on $M$ : $(\lambda, V) \leftarrow \text{eig}(M)$
Construct inverse root $\Sigma^{-1} \leftarrow \text{diag}\left(\sqrt{\frac{1}{\lambda_1}}, \ldots, \sqrt{\frac{1}{\lambda_R}}\right)$
Compute transformation matrix $B \leftarrow \sqrt{\lambda_R}\cdot V\Sigma^{-1}V^\intercal$
Compute task weight vector $\alpha \leftarrow Bw$
return $G\alpha$
```

C. Sampling Process

The sampling process is illustrated in the right part of Fig. 2. At each denoising timestep t, rather than predicting the noise

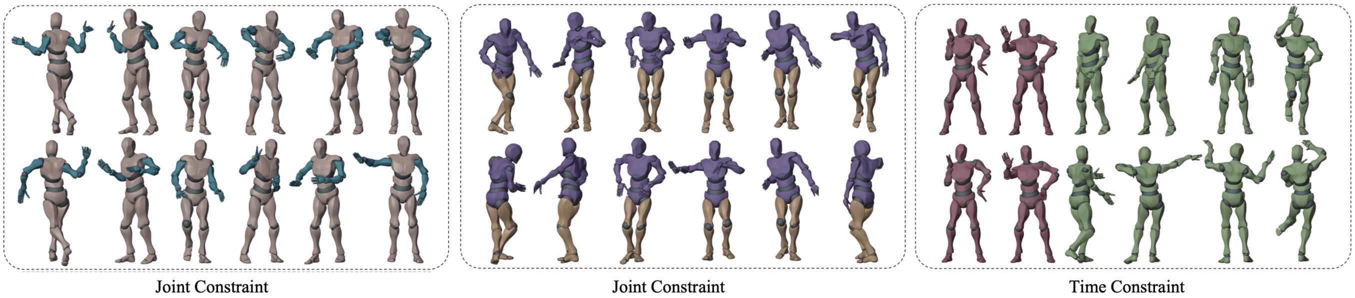

Fig. 5: GCDance supports joint-specific and temporally-specific generation under local constraints. In the left example, trunk joints are constrained and arm joints are generated. In the middle example, upper body joints are constrained and leg joints are generated. In the right example, the first one second is constrained and the final three seconds are generated.

term $\boldsymbol{d}_t$ for reconstruction like standard diffusion methods, our model directly estimates the dance pose $\hat{\boldsymbol{m}}$. This estimated pose is then diffused back to timestep $t-1$ per Equ. 14.

$$\boldsymbol{d}_{t-1} \sim q(\hat{\boldsymbol{m}}(\boldsymbol{d}_t, C_E, C_M), t-1) \tag{14}$$

This process is iterated until $t = 0$.

Editing Sampling. Building on [8], our approach employs diffusion inpainting during the sampling process to enable flexible, constraint-guided editing across joints and temporal segments, as depicted in Fig. 5. Given a partial reference motion $\boldsymbol{m}^{known}$ and a corresponding binary mask B indicating the constrained positions, the model denoises the sequence during sampling as follows.

$$\boldsymbol{d}_{t-1} := B \odot q(\boldsymbol{m}^{known}, t-1) + (1-B) \odot \boldsymbol{d}_{t-1} \tag{15}$$

where $\odot$ is the Hadamard product, which performs an element-wise substitution of known motion parts with noisy samples, guided by the specified constraint.

Given a reference motion $\boldsymbol{m}^{known} \in \mathbb{R}^{k\times 319}$ and a mask $B \in \{0,1\}^{k\times 319}$, with B marking hand-joint features as 0 and body-joint features as 1, the system generates a k-frame sequence in which body-joint movements are taken from the user reference, while coherent hand dance movements are synthesized for the hand joint regions. By offering flexible temporal and spatial control, the editing framework provides a robust capability for downstream applications, allowing the generation of dance sequences that adhere to diverse user-specified constraints.

Long-term Sampling. Building on editing capability, our model further supports the generation of long-term dance sequences with temporal consistency. Specifically, given a long music sequence, we split it into N sub-sequences of 4 seconds each. During the sampling process, GCDance enforces temporal continuity by aligning the first 2 seconds of each subsequence with the previous subsequence's last 2 seconds. To further maintain consistency between adjacent 2-second generated slices, we apply interpolation with linearly decaying weights to enhance performance. Through this approach, despite being trained on 4-second segments, our model supports the synthesis of arbitrarily long dances by applying temporal constraints across successive segments.

IV. Experimental Results & Analysis

This section presents the datasets, evaluation metrics, and experimental results. Additional implementation details are available in the supplementary material.

A. Dataset

The evaluation is conducted on the FineDance [14], AIST++ [2], and PopDanceSet [24] datasets. FineDance comprises 7.7 hours of music-dance pairs totaling $831,600$ frames at 30 frames per second across 16 genres. The mean dance duration is 152.3 seconds. The motion data is represented by a standard 52-joint 3D skeleton including finger joints. All methods are trained on the 183 music tracks in the training split. For evaluation, 18 songs from the test set are used to generate 270 dance clips, with the original corresponding dances serving as the ground truth. AIST++ contains $1,363$ 3D dance sequences paired with music, totaling 5.2 hours of motion data across 10 distinct genres at a frame rate of 60 FPS. PopDanceSet [24] contains 263 dance videos paired with 180 music tracks and covers 19 genres performed by 132 subjects, totaling 3.6 hours of motion data. Both AIST++ and PopDanceSet datasets employ a 24-joint skeleton representation based on the SMPL model [53]. For these two benchmarks, we strictly adhere to the official evaluation protocols for training and testing.

B. Evaluation Metrics

We evaluate our method based on four aspects: motion quality, motion diversity, motion-music correlation, and physical plausibility.

Motion Quality: We employ Fréchet Inception Distance (FID) [60] to evaluate motion quality. This metric calculates the feature space distance between the distributions of generated sequences and ground truth sequences to assess their dissimilarity.

Diversity: Following Bailando [4], we measure diversity based on the average Euclidean distance of kinetic features within the generated motion set.

Beat Synchronization: To assess the synchronization between motion transitions and music beats, we adopt the Beat Alignment Score (BAS) [4], which quantifies the alignment

TABLE I: Comparison on FineDance. We highlight the best result in **bold**, and the second-best in underline. ↓ denotes lower is better, ↑ denotes higher is better, and → denotes closer to the ground truth (GT) is better. $*$ denotes abnormally high Div values caused by discontinuous motions [2].

	Motion Quality		Motion Diversity		PFC↓	PBC→	BAS↑
	FID_hand↓	FID_body↓	Div_hand↑	Div_body↑			
GT	/	/	11.8156 ± 0.1314	10.1810 ± 0.1327	/	5.23 ± 0.16	0.2318 ± 0.0070
DanceRevolution [59]	219.52 ± 18.32	99.83 ± 7.79	1.85 ± 0.60	4.49 ± 0.25	6.81 ± 0.81	23.39 ± 2.03	0.2104 ± 0.0057
MNET [3]	195.56 ± 5.04	154.79 ± 2.80	6.79 ± 0.20	8.25 ± 0.39*	2.98 ± 0.11	12.21 ± 0.15	0.1792 ± 0.0014
Bailando [4]	55.60 ± 8.15	57.77 ± 6.01	6.40 ± 0.68	4.27 ± 0.43	0.34 ± 0.01	3.09 ± 0.06	0.2152 ± 0.0028
EDGE [8]	25.37 ± 3.24	51.56 ± 3.62	8.29 ± 0.30	5.88 ± 0.32	0.21 ± 0.03	7.78 ± 0.07	0.2171 ± 0.0056
FineNet [14]	26.88 ± 3.09	23.59 ± 3.56	8.30 ± 0.45	6.64 ± 0.28	**0.12 ± 0.01**	3.35 ± 0.11	0.2066 ± 0.0046
DGFM [10]	20.69 ± 3.52	24.63 ± 3.14	8.77 ± 0.41	6.77 ± 0.75	0.20 ± 0.01	4.23 ± 0.06	0.2153 ± 0.0054
LODGE [11]	18.36 ± 2.10	47.56 ± 1.37	8.57 ± 0.36	5.41 ± 0.27	0.13 ± 0.01	3.46 ± 0.06	**0.2327 ± 0.0050**
GCDance (Nash)	**17.69 ± 2.70**	22.90 ± 2.45	**9.47 ± 0.38**	6.39 ± 0.23	0.13 ± 0.01	**4.71 ± 0.06**	0.2238 ± 0.0056
GCDance (Aligned)	18.06 ± 3.12	**21.67 ± 2.41**	9.01 ± 0.47	**6.84 ± 0.75**	0.15 ± 0.01	4.54 ± 0.07	0.2205 ± 0.0041

by computing the average temporal distance between each kinematic beat and its nearest musical beat.

Physical Plausibility: We employ two complementary metrics to evaluate the physical realism of the generated motions. The Physical Foot Contact (PFC) metric [8] captures the consistency between the center-of-mass (COM) trajectory and foot-ground contact, while the Physical Body Contact (PBC) score [24] extends this concept by incorporating upper-body dynamics from the neck and hands.

TABLE II: Comparision on AIST++ and PopDanceSet.

	Method	Motion Quality		Motion Diversity		BAS↑
		FID_k↓	FID_m↓	Div_k↑	Div_m↑	
AIST++	FACT [2]	86.43	43.46	6.85	3.32	0.1607
	DanceNet [62]	69.18	25.49	2.86	2.85	0.1430
	Bailando [4]	28.16	**9.62**	**7.83**	6.34	0.2332
	DiffDance [9]	**24.09**	20.68	6.02	2.89	0.2418
	EDGE [8]	42.16	22.12	3.96	4.61	0.2334
	LODGE [11]	37.09	18.79	5.58	4.85	**0.2423**
	GCDance (Aligned)	35.91	19.19	5.07	5.70	0.2321
	GCDance (Nash)	30.93	18.25	5.22	**6.71**	0.2354
	Method	Motion Quality		Motion Diversity		BAS↑
		PFC↓	PBC→	Div_k↑	Div_m↑	
PopDanceSet	GroundTruth	1.2302	2.8485	6.4034	7.0289	0.330
	FACT [2]	7.5663	8.1007	3.7371	5.7843	0.405
	Bailando [4]	6.1762	5.9237	4.2253	5.5396	0.480
	EDGE [8]	5.9701	5.8535	3.6065	5.7350	0.475
	POPDG [24]	4.3697	5.3863	4.8641	6.0228	0.482
	GCDance (Aligned)	**2.3845**	4.5612	**5.2834**	6.0839	0.496
	GCDance (Nash)	3.3924	**4.3734**	4.4975	**6.7934**	**0.507**

C. Quantitative Results

In Table I, we compare our method with DanceRevolution [59], MNET [3], Bailando [4], EDGE [8], FineNet [14], DGFM [10] and LODGE [11]. Among these, only FineNet is originally trained on the FineDance dataset. For a fair comparison, we retrain the other methods on FineDance using their publicly available code and default training configurations. MNET is the only baseline that also incorporates genre information during generation. For every model, we produce ten sets of dance sequences by randomly drawing from the 270 clips in the test set. All sequences have length $T = 120$ frames, which corresponds to 4 seconds. We then evaluate their performance by reporting the mean and standard deviation of each metric across the ten sets.

The results show that, compared with EDGE, GCDance-Nash reduces FID_hand by 7.68 and FID_body by 28.66. Similarly, GCDance-Aligned reduces FID_hand by 7.31 and FID_body by 29.89. In terms of physical plausibility, GCDance-Nash and GCDance-Aligned achieve PFC scores of 0.13 ± 0.01 and 0.15 ± 0.01, which are close to the previous best FineNet, and both variants obtain the best PBC scores among all compared methods. Regarding the BAS score, our models are slightly lower than that of LODGE by 0.0111. Nevertheless, our model strikes a better balance between motion quality and diversity, leading to a more robust and generalizable performance. Additionally, it is worth noting that DanceRevolution and MNET achieve significantly higher FID scores, which we attribute to discontinuities in their generated motions. Furthermore, DanceRevolution often produces repeated or frozen frames, resulting in low diversity scores. In contrast, MNET often generates jittery motions with frequent spatial-temporal discontinuities. This instability introduces high-frequency noise that increases Euclidean distances in feature space, leading to abnormally high DIV values. We mark these values with an asterisk ($*$) to signify that these inflated metrics reflect poor motion quality rather than meaningful semantic diversity. Bailando demonstrates state-of-the-art performance on the 24-joint AIST++ dataset, as shown in Table II, but its performance degrades on the higher-resolution 52-joint FineDance dataset. This may be attributed to the model's design, which directly predicts joint positions rather than rotations [4], [61], potentially reducing accuracy when modeling more fine-grained skeletal structures.

To further validate the robustness of our framework across datasets, we retrain and evaluate our method on the AIST++ and PopDanceSet benchmarks, with results summarized in Table II. To ensure a fair and direct comparison with previous methods, we follow the official evaluation protocol of each benchmark, which leads to different metric sets across the datasets. For AIST++, despite the absence of genre annotations

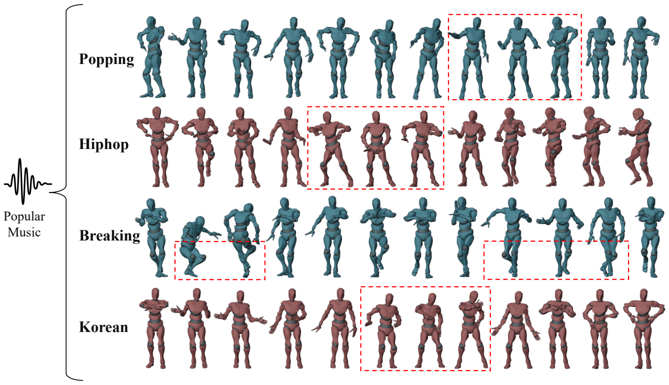

Fig. 6: **Same music, different popular dance.** Boxed hand, leg, and full-body poses highlight the salient stylistic features that distinguish each genre.

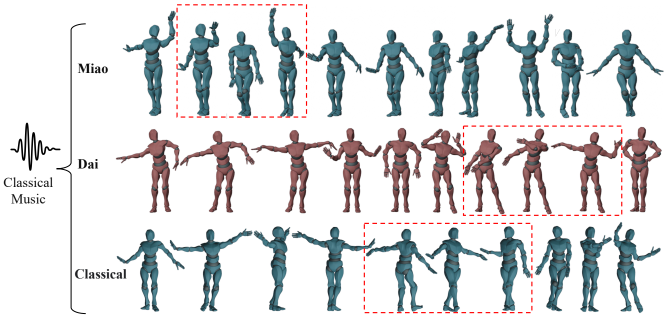

Fig. 7: **Same music, different classical dance.** Boxed hand, leg, and full-body poses highlight the salient stylistic features that distinguish each genre.

and genre-specific guidance, our model consistently surpasses EDGE across multiple metrics, demonstrating its ability to capture fundamental music–motion mappings. We further extend our evaluation to PopDanceSet to assess the framework under a different data distribution. As shown in Table II, GCDance achieves strong overall performance, improving physical plausibility and beat alignment while maintaining competitive motion diversity. Overall, these results across different benchmarks demonstrate the strong robustness and generalization of the proposed GCDance method.

D. Qualitative Results

To assess the controllability of our model in generating genre-specific dances, we fix the music content and vary the genre label only. For instance, using a single modern pop music segment as input, we generate four dances conditioned on Popping, Hiphop, Breaking, and Korean, so any stylistic differences are attributable to the label. The visualizations are shown in Fig. 6. We then apply the same protocol to a classical music segment, conditioning on Miao, Dai, and Classical dance to examine control under a different musical style. The corresponding results are presented in Fig. 7. In the first set of results, the generated Popping sequence features sharp hits with smooth transitional waves, Hip-Hop features abundant arm movements complemented by small rhythmic hops. Breaking emphasizes dynamic footwork, and Korean dance reproduces iconic K-pop elements. In the second set, Miao folk dance displays interlacing arm swings, Dai folk dance exhibits fluid and seamless movements, and Classical dance highlights broad arm gestures with graceful turns. These results demonstrate the controllability of GCDance in producing diverse stylistic performances from the same musical input. Additional demonstration videos are available on our project page.

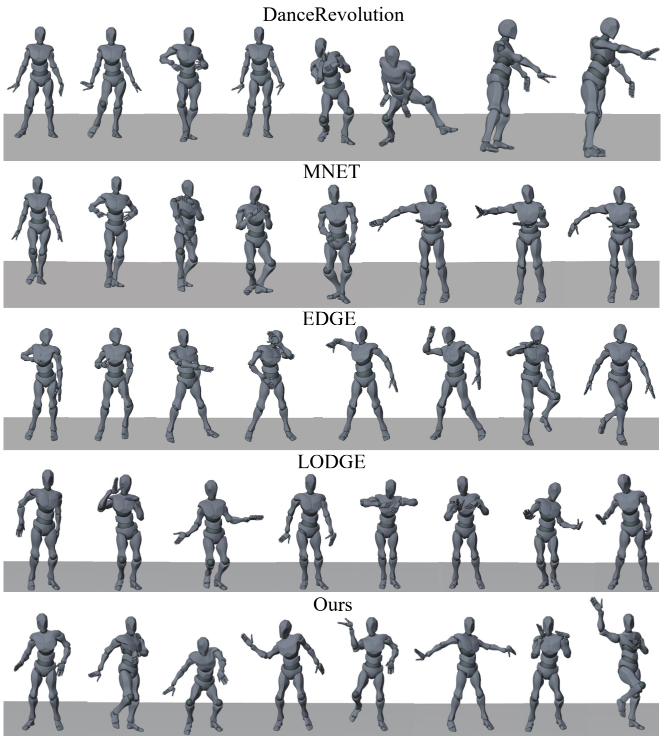

Fig. 8: A visual comparison with the SOTA methods.

TABLE III: Ablation study of each innovative component, including the use of Foundation Model (FM) features, Genre Classification Module (GCM), and different Multi-Task Learning (MTL) strategies.

System Components				Motion Quality		Motion Diversity		PFC ↓	BAS ↑
Base	**FM**	**GCM**	**Optimization**	FID$_h$ ↓	FID$_b$ ↓	Div$_h$ ↑	Div$_b$ ↑		
✓	-	-	Fixed	23.28	26.16	7.54	6.46	0.17	0.2155
✓	✓	-	Fixed	18.48	22.61	8.77	6.77	0.17	0.2188
✓	✓	✓	Fixed	**16.95**	30.31	8.92	6.36	0.15	0.2170
✓	✓	✓	Aligned-MTL	17.69	22.90	**9.47**	6.39	0.13	**0.2238**
✓	✓	✓	Nash-MTL	18.06	**21.67**	9.01	**6.84**	0.13	0.2205

Fig. 8 presents a qualitative comparison between our method and four baselines. DanceRevolution and MNET both suffer from motion stagnation after only a few seconds, reflecting poor temporal continuity and limited expressiveness. EDGE alleviates this freeze but introduces conspicuous artifacts, most notably unnatural hand trajectories and noticeable foot sliding. LODGE produces smoother kinematics yet offers reduced stylistic diversity. By contrast, our approach delivers motions with higher perceptual fidelity, richer stylistic variation, and coherence throughout the entire sequence.

E. Ablation Study

Table III presents a controlled, step-wise ablation designed to attribute the contribution of each innovative component of GCDance. We start from a diffusion baseline conditioned on

TABLE IV: Comparison of Dance Generation Quality Using Different Music Features

	Motion Quality		**Motion Diversity**		**PFC↓**	**PBC→**	**BAS↑**
	FID_h↓	FID_b↓	Div_h↑	Div_b↑			
GT	/	/	11.8156	10.1810	/	5.23	0.2318
CLAP [63]	29.64	27.52	8.11	6.10	0.23	3.43	0.2076
Wav2Vec2.0 [64]	21.78	34.65	8.61	6.32	0.20	3.93	0.2026
Jukebox [65]	23.02	32.26	7.41	6.38	0.24	3.35	0.2238
Wav2CLIP [13]	22.19	33.65	8.51	**8.85**	0.17	4.40	**0.2276**
STFT	23.70	25.84	7.60	6.85	0.16	3.86	0.2160
MFCC	27.63	33.74	8.42	8.50	0.19	3.81	0.2123
35-Feature Group* [14]	20.61	25.41	8.22	5.84	**0.15**	3.32	0.2028
Wav2CLIP+STFT (Ours)	**18.48**	**22.61**	**8.77**	6.77	0.17	**4.43**	0.2188

hand-crafted music features and trained with fixed loss weights following [8]. Adding music foundation model features improves motion quality and rhythm correlation, supporting the value of high-level semantic audio representations. Introducing the genre classification module strengthens controllability and reduces foot-contact artifacts, but it also exposes a clear trade-off where improvements in control come at the expense of some motion-quality and diversity indicators under fixed weighting. By replacing fixed weights with multi-task learning, the model re-balances these competing objectives, recovering the degraded quality metrics while further improving physical plausibility and beat alignment.

To quantify the effect of music features on dance quality, we compare generation results obtained with features extracted from multiple music foundation models, including CLAP [63], Wav2Vec2.0 [64], Jukebox [65], and Wav2CLIP [13], as well as hand-crafted features, including MFCC, STFT, and the 35-dimensional feature set provided by FineDance [14]. To fairly evaluate the impact of different music features, we use the backbone of our model without text classification and multi-task learning enhancements. This allows us to isolate the effect of music feature design from other factors. Table IV shows that our method, which incorporates music embeddings from a music foundation model and hand-crafted features, achieves the best overall performance. It demonstrates robust improvements in motion quality, diversity, and rhythm consistency over other music feature-based methods.

F. User study

We perform a user study with 20 participants at ANONYMIZED to measure the quality and controllability of the generated dance motions. For each method, 270 music–dance pairs are generated on the FineDance test set. From these, we randomly select the same 8 pairs across all the methods to ensure a fair comparison.

TABLE V: User Study on Generated Dance Samples.

Model	GCDance-Aligned	GCDance-Nash	FineDance	EDGE	Bailando
Wins	/	53.57%	63.26%	78.51%	89.28%
Control Score	46.87%	45.83%	/	/	/

TABLE VI: A comparison in model parameters and *per-instance* inference time.

Model	**# Parameters**	**Inference Time**
FACT [2]	120M	33.20s
Bailando [4]	152M	0.94s
EDGE [8]	50M	0.13s
FineNet [14]	94M	0.23s
LODGE [11]	236.8M	1.89s
GCDance (Ours)	88M	0.22s

For motion quality evaluation, participants are asked to rate each video on overall quality, smoothness, and synchronization with the music rhythm. As reported in Table V, our model consistently surpasses all baselines, achieving at least a 63.26% higher preference rate. For controllability evaluation, participants are presented with 8 pairs of dance videos sharing the same genre label, one generated with genre control and the other using the ground truth label. Participants are instructed to identify which video that better matches the given genre description. The results show that our generated dances are selected almost as frequently as the ground truth, confirming the strong genre controllability and consistency of GCDance.

G. Model Efficiency

Table VI compares model efficiency under a standardized inference setting. During inference, we report parameter counts and runtime for generating 4-second sequences. GCDance generates a sequence in 0.22 seconds with 87M parameters, offering a strong trade-off between model size and speed. It matches FineNet in runtime while using fewer parameters, and markedly outperforms Bailando, FACT, and LODGE in both efficiency and computational cost. Although EDGE attains the fastest runtime with a smaller model, GCDance provides a better overall trade-off between efficiency and motion generation quality.

V. Conclusion

In this paper, we introduced GCDance, a diffusion-based genre-specific dance generation framework driven by both music and textual prompts. By incorporating a genre classification module and leveraging features from a pretrained music foundation model, our method enabled precise and controllable synthesis of genre-consistent dance motions while preserving high motion quality and diversity. Furthermore, we used a multi-objective optimization strategy to balance the competing objectives, such as spatial accuracy, physical plausibility, and genre alignment, used for network training. Comprehensive evaluations on the FineDance, AIST++, and PopDanceSet datasets demonstrated the advantages of our method over the existing approaches both qualitatively and quantitatively.

However, the proposed GCDance method focuses on genre-level control of the generated dance sequences. It lacks the capability for fine-grained manipulation of specific motion attributes. In future work, we will enable text-driven control of individual joints with per-frame precision, allowing the textual prompt to vary across decoding steps.

VI. Acknowledgment

This work was partially supported by a joint PhD studentship from the China Scholarship Council (CSC) and the University of Surrey.

References

[1] G. Sun *et al.*, "DeepDance: music-to-dance motion choreography with adversarial learning," *IEEE Trans. Multimedia*, vol. 23, pp. 497–509, 2020.

[2] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI choreographer: Music conditioned 3D dance generation with AIST++," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13 401–13 412.

[3] J. Kim, H. Oh, S. Kim, H. Tong, and S. Lee, "A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3490–3500.

[4] L. Siyao *et al.*, "Bailando: 3D dance generation by actor-critic gpt with choreographic memory," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 11 050–11 059.

[5] K. Gong *et al.*, "TM2D: Bimodality driven 3D dance generation via music-text integration," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 9942–9952.

[6] G. Tevet *et al.*, "Human motion diffusion model," *arXiv preprint arXiv:2209.14916*, 2022.

[7] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 6840–6851, 2020.

[8] J. Tseng, R. Castellon, and K. Liu, "EDGE: Editable dance generation from music," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 448–458.

[9] Q. Qi *et al.*, "DiffDance: Cascaded human motion diffusion model for dance generation," in *Proc. of the 31st ACM Int. Conf. on Multimedia*, 2023, pp. 1374–1382.

[10] X. Liu, Z. Feng, D. Kanojia, and W. Wang, "DGFM: Full body dance generation driven by music foundation models," in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[11] R. Li *et al.*, "Lodge: A coarse to fine diffusion network for long dance generation guided by the characteristic dance primitives," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1524–1534.

[12] S. Lin, M. Zukerman, and H. Yan, "Music-driven choreography based on music feature clusters and dynamic programming," *IEEE Trans. Multimedia*, vol. 26, pp. 9330–9341, 2024.

[13] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2CLIP: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2022, pp. 4563–4567.

[14] R. Li *et al.*, "FineDance: A fine-grained choreography dataset for 3d full body dance generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 10 234–10 243.

[15] F. Ofli *et al.*, "An audio-driven dancing avatar," *J. on Multimodal User Interfaces*, vol. 2, pp. 93–103, 2008.

[16] S. Fukayama and M. Goto, "Music content driven automated choreography with beat-wise motion connectivity constraints." Citeseer, 2015, pp. 177–183.

[17] D. Holden, J. Saito, and T. Komura, "A deep learning framework for character motion synthesis and editing," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 1–11, 2016.

[18] H. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, "Action-agnostic human pose forecasting," in *2019 IEEE Winter Conf. on Appl. of Comput. Vis. (WACV)*. IEEE, 2019, pp. 1423–1432.

[19] X. Du, R. Vasudevan, and M. Johnson-Roberson, "Bio-LSTM: A biomechanically inspired recurrent neural network for 3-D pedestrian pose and gait prediction," *IEEE Robot. Autom. Lett.*, vol. 4, no. 2, pp. 1501–1508, 2019.

[20] D. Fan, L. Wan, W. Xu, and S. Wang, "A bi-directional attention guided cross-modal network for music based dance generation," *Comput. and Electrical Engineering*, vol. 103, p. 108310, 2022.

[21] Y. Huang *et al.*, "Genre-conditioned long-term 3D dance generation driven by music," in *ICASSP 2022-2022 IEEE Int. Conf. on Acoust., Speech and Signal Process.*, 2022, pp. 4858–4862.

[22] B. Li, Y. Zhao, S. Zhelun, and L. Sheng, "DanceFormer: Music conditioned 3D dance generation with parametric motion transformer," in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 1272–1279.

[23] W. Zhuang *et al.*, "Music2Dance: Dancenet for music-driven dance generation," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 2, pp. 1–21, 2022.

[24] Z. Luo, M. Ren, X. Hu, Y. Huang, and L. Yao, "POPDG: Popular 3D dance generation with PopDanceSet," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 26 984–26 993.

[25] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *Int. Conf. Mach. Learn.* PMLR, 2021, pp. 8162–8171.

[26] F.-A. Croitoru, V. Hondru, R. T. Ionescu, and M. Shah, "Diffusion models in vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 9, pp. 10 850–10 869, 2023.

[27] N. Ruiz *et al.*, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 22 500–22 510.

[28] Y. Xu, X. Xu, H. Gao, and F. Xiao, "SGDM: an adaptive style-guided diffusion model for personalized text to image generation," *IEEE Trans. Multimedia*, vol. 26, pp. 9804–9813, 2024.

[29] H. Liu *et al.*, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 2871–2883, 2024.

[30] H. Liu *et al.*, "AudioLDM: Text-to-audio generation with latent diffusion models," *Proc. IEEE Int. Conf. on Mach. Learn*, pp. 21 450–21 474, 2023.

[31] C. Zhang, Y. Ren, K. Zhang, and S. Yan, "SDMuse: Stochastic differential music editing and generation via hybrid representation," *IEEE Trans. Multimedia*, vol. 26, pp. 1681–1689, 2023.

[32] J. Lovelace, V. Kishore, C. Wan, E. Shekhtman, and K. Q. Weinberger, "Latent diffusion for language generation," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[33] Z. He, T. Sun, K. Wang, X. Huang, and X. Qiu, "DiffusionBERT: Improving generative masked language models with diffusion models," *arXiv preprint arXiv:2211.15029*, 2022.

[34] H. Chung, B. Sim, D. Ryu, and J. C. Ye, "Improving diffusion models for inverse problems using manifold constraints," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 25 683–25 696, 2022.

[35] M. Zhao, W. Wang, T. Chen, R. Zhang, and R. Li, "TA2V: Text-audio guided video generation," *IEEE Trans. Multimedia*, vol. 26, pp. 7250–7264, 2024.

[36] A. Nichol *et al.*, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[37] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10 684–10 695.

[38] O. Avrahami, D. Lischinski, and O. Fried, "Blended diffusion for text-driven editing of natural images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18 208–18 218.

[39] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[40] K. Karunratanakul, K. Preechakul, S. Suwajanakorn, and S. Tang, "Guided motion diffusion for controllable human motion synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 2151–2162.

[41] S. Alexanderson, R. Nagy, J. Beskow, and G. E. Henter, "Listen, denoise, action! audio-driven motion synthesis with diffusion models," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–20, 2023.

[42] T. Standley *et al.*, "Which tasks should be learned together in multi-task learning?" in *Proc. of the 37th Int. Conf. on Mach. Learn.*, 2020.

[43] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1871–1880.

[44] L. Liu *et al.*, "Towards impartial multi-task learning," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.

[45] A. Navon *et al.*, "Multi-task learning as a bargaining game," in *Proc. of the 39th Int. Conf. on Mach. Learn.*, 2022, pp. 16 428–16 446.

[46] J. Nash, "Two-person cooperative games," *Econometrica*, vol. 21, no. 1, pp. 128–140, 1953.

[47] D. Senushkin, N. Patakin, A. Kuznetsov, and A. Konushin, "Independent component alignment for multi-task learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 20 083–20 093.

[48] I. Achituve, I. Diamant, A. Netzer, G. Chechik, and E. Fetaya, "Bayesian uncertainty for gradient aggregation in multi-task learning," *arXiv preprint arXiv:2402.04005*, 2024.

[49] G. Tu *et al.*, "A multi-task neural approach for emotion attribution, classification, and summarization," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 148–159, 2019.

[50] H. Zhang, S. Qian, Q. Fang, and C. Xu, "Multi-modal meta multi-task learning for social media rumor detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1449–1459, 2021.

[51] S. Deoghare *et al.*, "A multi-task learning framework for quality estimation," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 9191–9205.

[52] S. Qian, C. Orăsan, D. Kanojia, and F. d. Carmo, “A multi-task learning framework for evaluating machine translation of emotion-loaded user-generated content,” *arXiv preprint arXiv:2410.03277*, 2024.

[53] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, “SMPL: A skinned multi-person linear model,” in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2023, pp. 851–866.

[54] M. K. Gourisaria, R. Agrawal, M. Sahni, and P. K. Singh, “Comparative analysis of audio classification with mfcc and stft features using machine learning techniques,” *Discover Internet of Things*, vol. 4, no. 1, p. 1, 2024.

[55] B. McFee *et al.*, “librosa: Audio and music signal analysis in python,” in *SciPy*, 2015, pp. 18–24.

[56] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *Int. J. of Comput. Vis.*, vol. 130, no. 9, pp. 2337–2348, 2022.

[57] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. of the AAAI Conf. on Artificial Intelligence*, vol. 32, no. 1, 2018.

[58] C. Saharia *et al.*, “Photorealistic text-to-image diffusion models with deep language understanding,” *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 36 479–36 494, 2022.

[59] R. Huang *et al.*, “Dance revolution: Long-term dance generation with music via curriculum learning,” *arXiv preprint arXiv:2006.06119*, 2020.

[60] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[61] S. Yang, Z. Yang, and Z. Wang, “LongDanceDiff: Long-term dance generation with conditional diffusion model,” *arXiv preprint arXiv:2308.11945*, 2023.

[62] W. Zhuang, C. Wang, S. Xia, J. Chai, and Y. Wang, “Music2Dance: Music-driven dance generation using wavenet,” *arXiv preprint arXiv:2002.03761*, 2020.

[63] Y. Wu *et al.*, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP 2023-2023 IEEE Int. Conf. on Acoust., Speech and Signal Process.* IEEE, 2023, pp. 1–5.

[64] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 12 449–12 460, 2020.

[65] P. Dhariwal *et al.*, “Jukebox: A generative model for music,” *arXiv preprint arXiv:2005.00341*, 2020.

Xinran Liu received the B.S. degree in Computer Science and Technology from Shandong Agricultural University, Tai’an, China, in 2020, and the M.Eng. degree in Computer Technology from the University of Science and Technology Beijing, Beijing, China, in 2023. She is currently pursuing the Ph.D. degree with the School of Computer Science and Electronic Engineering, University of Surrey, Guildford, U.K. Her research interests include multimodal learning and generative modeling, with a focus on conditional human motion generation.

Xu Dong received the B.Sc. and M.Sc. degrees from the University of Sheffield and Queen Mary University of London, U.K. He is currently pursuing the Joint Dual-Degree Ph.D. degree with the University of Surrey, U.K., and the University of Wollongong, Australia. His research interests include computer vision and video understanding. He has published papers in several conferences and journals, including IJCV, BMVC, IEEE FG, etc.

Shenbin Qian is a postdoctoral researcher in the Language Technology Group at the University of Oslo, funded by the Marie Skłodowska-Curie Actions programme DSTrain. He also contributes to the Digital Europe consortium OpenEuroLLM: Foundation Models for Transparent AI in Europe. He received his PhD from the University of Surrey, where his research spanned multiple areas of machine learning, including machine translation evaluation, information retrieval, text-to-image generation, and video understanding. His current research focuses on advancing the understanding and development of large language models, with particular emphasis on explainability and multilingual evaluation.

Diptesh Kanojia is a senior lecturer jointly at Surrey Institute for People-Centred AI (SI-PAI), and Computer Science Research Centre, University of Surrey. His research focuses on scalable approaches using foundation models for safe human-machine and human-human interaction within social and personal digital spaces. With a People-Centred focus on Responsible and Inclusive AI, his work contributes to low-resource languages and addressing societal challenges like digital accessibility, online toxicity, and misinformation. Broadly, his interests lie in modelling information from linguistics, audio-visual signals (Multimodality), social phenomena (SocialNLP), and cognitive behaviour (CognitiveNLP), to improve machine understanding. He received his B.Tech. degree in 2013, and his joint PhD from IIT Bombay and Monash University in 2021.

Wenwu Wang (M’02-SM’11-F’26) received the B.Sc., M.E., and the Ph.D. degrees in automation, from Harbin Engineering University, China, in 1997, 2000, and 2002, respectively. He then worked with King’s College London, Cardiff University, Tao Group Ltd. (now Antix Labs Ltd.), and Creative Labs, before joining University of Surrey, U.K., in 2007, where he is currently a Professor in Signal Processing and Machine Learning. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His research interests include signal processing, machine learning and perception, machine audition (listening). He has (co)-authored over 400 papers in these areas. His works have been recognized with various awards, including the Audio Engineering Society Best Technical Paper Award (2025), IEEE Signal Processing Society Young Author Best Paper Award (2022), DCASE Judge’s Award (2020, 2023, and 2024), and DCASE Reproducible System Award (2019 and 2020). He has been an invited Keynote or Plenary Speaker on more than 20 international conferences and workshops.

Zhenhua Feng (S’13-M’16-SM’22) received the Ph.D. degree from the Centre for Vision, Speech and Signal Processing, University of Surrey, UK, in 2016. He then worked as Research Fellow, Senior Research Fellow, Lecturer, and Senior Lecturer at the University of Surrey from 2016 to 2024. He is currently a full Professor at the School of Artificial Intelligence and Computer Science, Jiangnan University, China. His research interests include computer vision and machine learning. He has published more than 100 scientific papers in top-tier venues. He received the 2024 ICPR Best Scientific Paper Award, the 2017 European Biometrics Industry Award from the European Association for Biometrics (EAB), the AMDO 2018 Best Paper Award for Commercial Applications, etc. He currently serves as an Associate Editor for IEEE TNNLS and Complex & Intelligent Systems, and Guest Editor for IEEE Trans. on Games. He also served as the Guest Editor for IJCV (2023), Program Chair for BMVC 2022, Area Chair for BMVC 2021-2025, and CVMP 2022/23.