# Controllable Group Choreography using Contrastive Diffusion

NHAT LE, AIOZ, Singapore
TUONG DO, AIOZ, Singapore
KHOA DO, VNUHCM-University of Science, Vietnam
HIEN NGUYEN, AIOZ, Singapore
ERMAN TJIPUTRA, AIOZ, Singapore
QUANG D.TRAN, AIOZ, Singapore
ANH NGUYEN, University of Liverpool, United Kingdom

Fig. 1. We present a contrastive diffusion method that controls the consistency (top row) and diversity (second row) in group choreography.

Music-driven group choreography poses a considerable challenge but holds significant potential for a wide range of industrial applications. The ability to generate synchronized and visually appealing group dance motions that are aligned with music opens up opportunities in many fields such as entertainment, advertising, and virtual performances. However, most of the recent works are not able to generate high-fidelity long-term motions, or fail to enable controllable experience. In this work, we aim to address the demand for high-quality and customizable group dance generation by effectively governing the consistency and diversity of group choreographies. In particular, we utilize a diffusion-based generative approach to enable the synthesis of flexible number of dancers and long-term group dances, while ensuring coherence to the input music. Ultimately, we introduce a Group Contrastive Diffusion (GCD) strategy to enhance the connection between dancers and their group, presenting the ability to control the consistency or diversity level of the synthesized group animation via the classifier-guidance sampling technique. Through intensive experiments and evaluation, we demonstrate the effectiveness of our approach in producing visually captivating and consistent group dance motions. The experimental results show the capability of our method to achieve the desired levels of consistency and diversity, while maintaining the overall quality of the generated group choreography.

CCS Concepts: • **Computing methodologies** → **Animation/Simulation**; Machine Learning Approaches; Methods & Applications.

Additional Key Words and Phrases: Group Choreography Animation, Group Motion Synthesis, Machine Learning, Diffusion Models.

**ACM Reference Format:**
Nhat Le, Tuong Do, Khoa Do, Hien Nguyen, Erman Tjiputra, Quang D.Tran, and Anh Nguyen. 2023. Controllable Group Choreography using Contrastive Diffusion. *ACM Trans. Graph.* 42, 6, Article 224 (December 2023), 14 pages. https://doi.org/10.1145/3618356

## 1 INTRODUCTION

With the widespread presence of digital social media platforms, the act of creating and editing dance videos has gained immense popularity among social communities. This surge in interest has resulted in the daily production and watching of millions of dancing videos across online platforms [Fink et al. 2021; Kico et al. 2018]. Recently, researchers from computer vision, computer graphics, and machine learning communities have devoted considerable attention to developing techniques that can generate natural dance movements from music [Bisig 2022]. These advancements have far-reaching implications and find applications in various domains, such as animation [Li et al. 2021b], the creation of virtual idols [Perez et al. 2021; Pham et al. 2023], the development of virtual meta-verse [Lee et al. 2021], and dance education [Alaoui et al. 2014; Shi 2021; Soga et al. 2005]. These techniques empower artists, animators, and educators alike, providing them with powerful tools to enhance their creative

Authors' addresses: Nhat Le, nhat.minh.le@aioz.io, AIOZ, Singapore; Tuong Do, tuong.khanh-long.do@aioz.io, AIOZ, Singapore; Khoa Do, 19110348@student.hcmus.edu.vn, VNUHCM-University of Science, Vietnam; Hien Nguyen, hien.nguyen@aioz.io, AIOZ, Singapore; Erman Tjiputra, erman.tjiputra@aioz.io, AIOZ, Singapore; Quang D.Tran, quang.tran@aioz.io, AIOZ, Singapore; Anh Nguyen, anh.nguyen@liverpool.ac.uk, University of Liverpool, United Kingdom.


0730-0301/2023/12-ART224 $15.00
https://doi.org/10.1145/3618356

Our project page is available at: https://aioz-ai.github.io/GCD/

endeavors and enrich the dance experience for both performers and audiences.

While significant progress has been made in generating dancing motions for single dancer [Ferreira et al. 2021; Huang et al. 2020; Kim et al. 2022b; Li et al. 2021b; Perez et al. 2021; Siyao et al. 2022; Tseng et al. 2023], the task of producing cohesive and expressive choreography for a group of dancers has received limited attention [Le et al. 2023]. The generation of synchronized group dance motions that are both realistic and aligned with music remains a challenging problem in the field of computer animation and motion synthesis [Chen et al. 2021; Yalta et al. 2019]. This is primarily due to the complex relationship between music and human motion, the diverse range of motions required for group performances, and the insufficient of a suitable dataset [Le et al. 2023]. At present, AIOZ-GDance [Le et al. 2023] stands as the most recent extensive dataset available to facilitate the task of generating group choreography. Besides, while current algorithms can generate individual movements and choreographic sequences, ensuring that these elements align seamlessly with the overall group performance is also paramount [Tsuchida et al. 2019].

Different from solo dance, group dance involves coordination and interaction between dancers, making it crucial and challenging to establish correlations between motion series within a group [Le et al. 2023]. Besides, group dance can involve complex and diverse choreographies among participating dancers while still maintaining a semantic relationship between the motion and input music. Exploring the ***consistency*** and ***diversity*** between the movements of dancers of the synthesized group choreography is of vital importance to create a natural and expressive performance. The ability to control the consistency and diversity in group dance generation holds great potential across various applications [Bisig 2022]. One such application is in the realm of entertainment and performance. Choreographers and creative teams can leverage this control ability to design captivating group dance routines that seamlessly blend synchronized movements with moments of individual expression. Second, in the context of animation and virtual metaverses, the control over consistency and diversity allows for the creation of visually stunning and immersive virtual dance performances. By balancing the synchronization of dancers' movements, while also introducing variations and unique flourishes, the generated group dances can captivate audiences and evoke a sense of realism and authenticity. Last but not least, in dance education and training, the ability to regulate consistency and diversity in group dance generation can be invaluable. It enables instructors to provide students with a diverse range of generated dance routines and samples that challenge their abilities, promote collaboration, and foster creativity. By dynamically adjusting the level of consistency and diversity, educators can cater to the unique need and skill level of each individual dancer, creating more inclusive instructions and enriching the learning environment [Phillips et al. 2009]. Although plenty of applications can be listed, due to some limitations of data establishment [Le et al. 2023], investigating the consistency and diversity in group choreography has not been carefully explored.

In this paper, our goal is to develop a controllable technique for group dance generation. We present a **G**roup **C**ontrastive **D**iffusion (GCD) strategy that learns an encoder to capture the key targets between group dance movements. Diffusion modeling provides a flexible framework for manipulating the dance distribution, which allows us to modulate the degree of diversity and consistency in the generated dances. By using denoising diffusion probabilistic model [Ho et al. 2020] as a key technique, we can effectively control the trade-off between diversity and consistency during the group dance generation, thanks to the guided sampling process. With this approach, we can guide the generation process toward a desired balance between diversity and consistency levels. Moreover, incorporating the encoder, which learns the association between the dancers and their group, can help to maintain the generated dance moves so that they are consistent with a specific dance style, music genre, or any long-term chorus. We empirically show that this approach has the potential to enhance the quality and naturalness of generated group dance performances, making it more appealing for various applications.

To summarize, our key contributions are as follows:

- We introduce contrastive diffusion, the first denoising diffusion approach for music-driven group choreography. Our model is able to generate high-fidelity and diverse group dancing motions that are aligned with the input music.
- We develop a method to trade-off between the consistency and diversity of generated group motions. Our framework allows users to control and generate different outputs from a single piece of input music.
- Extensive experiments along with user study evaluations demonstrate state-of-the-art performance of our model in synthesizing group choreography animation, as well as creating long dance motion sequences while maintaining the coherency among dancers.

## 2 RELATED WORK

### 2.1 Music-driven Choreography

Creating natural and authentic human choreography from music is a complex task [Joshi and Chakrabarty 2021]. One commonly employed technique involves using a motion graph derived from a vast motion database to generate new motions [Kovar et al. 2002]. This involves combining various motion segments and optimizing transition costs along the graph path. Alternatively, there are other methods that incorporate music-motion similarity matching constraints to ensure consistency between the motion and the accompanying music [Kim et al. 2003; Safonova and Hodgins 2007]. Previous studies have extensively explored these methodologies [Fan et al. 2011; Kim et al. 2006; Lee et al. 2013; Shiratori et al. 2006]. However, most of these approaches relied on heuristic algorithms to stitch together pre-existing dance segments sourced from a limited music-dance database [Fan et al. 2011]. While these methods are successful in generating extended and realistic dance sequences, they face limitations when trying to create entirely novel dance fragments [Ofli et al. 2011].

In recent years, several signs of progress have been made in the field of music-to-dance motion generation using Convolutional Network (CNN) [Ahn et al. 2020; Chan et al. 2019; Sun et al. 2020; Ye et al. 2020; Yin et al. 2022; Zhuang et al. 2022], Recurrent Network (RNN) [Alemi et al. 2017; Huang et al. 2020; Sun et al. 2020; Tang

et al. 2018; Yalta et al. 2019], Graph Neural Network (GNN) [Au et al. 2022; Ferreira et al. 2021; Ren et al. 2020; Zhou and Luo 2022], Generative Adversarial Network (GAN) [Lee et al. 2019; Sun et al. 2020], or Transformer [Kim et al. 2022b; Li et al. 2022a,b, 2021b; Perez et al. 2021; Siyao et al. 2022]. Typically, these methods rely on multiple inputs such as the current music and a brief history of past dance movements to predict the future sequence of human poses. Recently, Gong et al. [2023] propose an interesting task of generating dance by simultaneously utilizing both music and text instruction. A music-text feature fusion module was designed to fuse the inputs into a motion decoder to generate dance conditioned on both music and text. However, although these methods have the potential to produce natural and realistic dancing motion, they are often unable to create synchronized and harmonious movements between multiple dancers [Le et al. 2023]. Ensuring coordination and synchronization between dancers is a complicated problem, as it involves not only individual pose predictions but also the seamless integration of these poses within the context of a group. Achieving synchronized and harmonious group movements requires considering spatial and temporal relationships among dancers, their interactions, and the overall choreographic structure [Tsuchida et al. 2019]. Thus, further advances in the field are considered to address these challenges, including works that use deep learning approaches such as Variational Autoencoder (VAE) [Hong et al. 2022; Li et al. 2021a], GAN [Zhu et al. 2022], and Normalising Flow [Perez et al. 2021]. Zhou et al. [2019b] explore the use of VAE to combine motion data with style embeddings so as to generate diverse and stylistically consistent dance movements. Meanwhile, Huang and Liu [2021] introduce a conditional GAN-based approach for generating new dance motions. Perez et al. [2021] combine a multimodal transformer encoder with a normalising-flow-based decoder to estimate a probability distribution encompassing the potential succeeding poses. Unfortunately, most of these networks are limited by their ability to model long-term dance sequences (e.g., over 8 seconds) as the generated sequence may freeze or drift towards the end of the music [Sun et al. 2022]. Feng et al. [2023] learn the dance movements using unpaired data with music style and motion style exemplars of the same style. To facilitate long-term generation, they apply a motion repeat constraint to predict future frames by attending to the historical motions. Nevertheless, this would limit the flexibility of the model by forcing it to always look into the past. Aristidou et al. [2022] use the motion motifs (clusters of similar short motion sequences) and motion signatures [Aristidou et al. 2018] to guide the dance synthesis to preserve the global consistency following a specific dance style. Consistency in *single dance* (as mentioned in [Aristidou et al. 2022]) is related to temporal information of a motion sequence itself, whereas diversity and consistency in *group dance* paradigm are factors between motions of two or more dancers within a period.

## 2.2 Group Choreography

Group choreography and its related problem, multi-person motion prediction, have been an active research area with numerous studies addressing the challenges of predicting the behaviors of multiple individuals [Aliakbarian et al. 2020; Arikan and Forsyth 2002; Chen et al. 2020; Guo et al. 2022; Khaire and Kumar 2022; Kiciroglu et al. 2022; Kim et al. 2022a; Mehta et al. 2018; Song et al. 2022; Stergiou and Poppe 2019]. One approach from Alahi et al. [2014] utilizes a Markov chain model to jointly analyze the trajectories of several pedestrians and predict their destinations in a given scene. Another method presented in [Adeli et al. 2020] integrates social interactions and the visual context of the environment to forecast the future motion of multiple individuals. Multi-Range Transformers, introduced by Wang et al. [2021], has the capability to predict the movements of groups with more than ten people engaging in social interactions. Recently, Le et al. [2023] develop a novel approach that utilizes input music sequences and a set of 3D positions of dancers to generate multiple choreographies with group coherency. Wang et al. [2022] present a collaboration system to determine which period the dancers should perform dancing with each other and then produce the corresponding motion sequence for each dancer. These aforementioned methods leverage various techniques to capture social interactions [Arikan and Forsyth 2002; Willis et al. 2004], spatial dependencies [Aliakbarian et al. 2020; Mehta et al. 2018], and temporal dynamics [Chen et al. 2020; Stergiou and Poppe 2019], generally aiming to predict accurate and socially plausible future motions for multiple individuals in different scenarios. However, despite the notable advancements achieved, there remains a demand for further investigation of the correlation between the consistency and diversity of motions within group context [Le et al. 2023]. A deeper understanding of how to attain the optimal balance between consistency and diversity holds the potential to unlock new possibilities for creating group choreographies that benefit the users in many circumstances.

## 2.3 Diffusion for Music-driven Choreography

Recently, diffusion-based approaches have shown remarkable results on several generative tasks [Yang et al. 2022] ranging from image generation [Dhariwal and Nichol 2021; Nichol et al. 2022a; Ramesh et al. 2022; Rombach et al. 2022; Saharia et al. 2022], audio synthesis [Kong et al. 2020; Popov et al. 2021], pose estimation [Nguyen et al. 2023], natural language generation [Nichol and Dhariwal 2021], and motion synthesis [Alexanderson et al. 2023; Dabral et al. 2023; Ren et al. 2023; Tevet et al. 2023], to point cloud generation [Luo and Hu 2021; Nichol et al. 2022b], 3D object synthesis [Poole et al. 2022; Seo et al. 2023; Xiang et al. 2023], and scene creation [Huang et al. 2023; Sharp et al. 2022; Vuong et al. 2023; Zeng et al. 2022]. Diffusion models have shown that they can achieve high mode coverage, unlike GANs, while still maintaining high sample quality [Ulhaq et al. 2022; Yang et al. 2022]. This ability makes them an ideal method for the music-to-dance generation task. Dabral et al. [2023] introduce a denoising diffusion-based framework that enables the generation of extended, realistic, and semantically faithful human motion sequences by considering diverse conditioning contexts (e.g., text or music). Tseng et al. [2023] present an editable dance generation model (EDGE), which exploits the capability of a transformer-based diffusion architecture and a strong music feature extractor, to provide flexible editing capabilities for dance applications. Most existing diffusion-based approaches for human motion/dance synthesis only focus on generating motion sequences for *a single character*, conditioned on information such as text [Tevet et al. 2023; Zhang et al.

2022], audio [Alexanderson et al. 2023; Tseng et al. 2023], or both audio and text [Dabral et al. 2023; Zhou and Wang 2023]. Different from these prior works, we aim to create *group of dancing motions* from music, which includes coordinating multiple characters, avoiding collisions, and maintaining coherence between them. In addition to the vanilla diffusion loss term used for training in previous works, our method employs a contrastive learning strategy that directly influences the training of the diffusion reverse process, enhancing the association within the dance group. In terms of controllable generation, while the work of Tevet et al. [2023] can edit individual motion sequence using text prompts or Alexanderson et al. [2023] interpolates different motion styles using classifier-free guidance [Ho and Salimans 2022], our approach provides the means to control the trade-off between consistency and diversity of group dance through the learned contrastive encoder. Recently, Chopin et al. [2023] propose BiGraphDiff, a diffusion approach based on bipartite graph architecture for text-driven human motion interactions between two persons. Concurently, Shafir et al. [2023] train a small communication block between two pre-trained Motion Diffusion Models [Tevet et al. 2023] to coordinate between two instances for two-person motion generation from text prompts. Limited by the architectural designs, these methods can only synthesize motion for only two persons, while our model is capable of generating group dancing motion with flexible number of dancers.

A prominent issue of the diffusion approach for motion synthesis is that although it is highly effective in generating diverse samples, injecting a large amount of noise during the sampling process can lead to inconsistent results. This issue is particularly problematic for group dance paradigm. Therefore, our desideratum is to design a group dance generation model with the ability to address both diversity and consistency problem. Typically, the work in [Le et al. 2023] mainly addresses the consistency through a cross-entity attention mechanism, but it is not entirely effective as diversity is overlooked due to the deterministic nature of their main training process. Different from them, we devise a contrastive diffusion approach to tackle these issues altogether. Our model is not only able to create numerous distinctive dancing motions while preserving their coherency, but it also has the flexibility to allow the user to freely control the diversity or consistency level.

## 3 METHODOLOGY

### 3.1 Background

Given an input music sequence $\{a_1, a_t, ..., a_T\}$ with $t = \{1, ..., T\}$ indicates the index of the music frames, our goal is to generate the group motion sequences of $N$ dancers: $\{x_1^1, ..., x_T^1; ...; x_1^N, ..., x_T^N\}$ where $x_t^i$ is the pose of $i$-th dancer at frame $t$. We represent dances as sequences of poses in the 24-joint of the SMPL model [Loper et al. 2015], using the 6D continuous rotation [Zhou et al. 2019a] for every joint, along with a single 3D root translation. This rotation representation ensures the uniqueness and continuity of the rotation vector, which is more beneficial to the training of deep neural networks. We tackle the group dance generation task by using a diffusion-based framework to synthesize the motions from a random noise distribution, given the music conditioning. Thanks to the sampling process of the diffusion model, we can effectively control the consistency and diversity in the generated sequences.

**Forward Process of Diffusion Model.** Given an original sample from the real data distribution $x_0 \sim q(x_0)$, following [Ho et al. 2020], the forward diffusion process is defined as a Markov process that gradually adds Gaussian noise to the data under a pre-defined noise schedule up to $M$ steps.

$$q(x_m|x_{m-1}) = \mathcal{N}(x_m; \sqrt{1-\beta_m}x_{m-1}, \beta_m I), \forall m \in \{1, 2, ..., M\} \quad (1)$$

If the noise variance schedule $\beta_m$ is small and the number of diffusion step $M$ is large enough, the distribution $q(x_M)$ at the end of the process is well-approximated by a standard normal distribution $\mathcal{N}(0, I)$, which is easy to sample from. Thanks to the nice property of the forward diffusion, we can directly obtain the noised sample at any arbitrary step $m$ without traversing through the whole chain:

$$q(x_m|x_0) = \mathcal{N}(x_m; \sqrt{\bar{\alpha}_m}x_0, (1-\bar{\alpha}_m)I), \quad (2)$$

$$x_m = \sqrt{\bar{\alpha}_m}x_0 + \sqrt{1-\bar{\alpha}_m}\epsilon, \epsilon \sim \mathcal{N}(0, I) \quad (3)$$

where $\alpha_m = 1 - \beta_m$ and $\bar{\alpha}_m = \prod_{s=0}^{m} \alpha_s$.

**Reverse Process.** By additionally conditioning on $x_0$, the posterior of the reverse process is tractable and becomes a Gaussian distribution:

$$q(x_{m-1}|x_m, x_0) = \mathcal{N}(x_{m-1}; \tilde{\mu}_m, \tilde{\beta}_m I), \quad (4)$$

where $\tilde{\mu}_m$ and $\tilde{\beta}_m$ are the posterior mean and variance that depend on both $x_m$ and $x_0$, respectively. We refer the readers to [Ho et al. 2020] for a detailed derivation of the posterior mean and variance. To obtain a sample from the original data distribution, we start by sampling from the noise distribution $q(x_M)$ and then gradually remove the noise until we reach $x_0$, following the reverse process. Therefore, our goal is to train a neural network to approximate the posterior $q(x_{m-1}|x_m)$ of the reverse process as:

$$p_\theta(x_{m-1}|x_m) = \mathcal{N}(x_{m-1}; \mu_\theta(x_m, m), \Sigma_\theta(x_m, m)) \quad (5)$$

We follow [Ho et al. 2020] to model only the mean $\mu_\theta(x_m, m)$ of the reverse distribution while keeping the variance $\Sigma_\theta(x_m, m)$ fixed according to the noise schedule. However, instead of predicting the noise $\epsilon_m$ at any arbitrary step $m$ as in their approach, we train the network to learn to predict the original noiseless signal $x_0$. The sample at the previous step $m-1$ can be obtained by noising back the predicted $x_0$ through Equation 3. For conditional generation setting, the network is additionally conditioned with the conditioning signal $c$ as $x_0 \approx \mathcal{G}_\theta(x_m, m, c)$ with model parameters $\theta$.

### 3.2 Group Diffusion Denoising Network

Our model architecture is illustrated in Figure 2. We utilize a transformer based architecture to generate the whole sequence in one go. Compared with recent auto-regressive approach [Le et al. 2023], our method does not suffer from the error accumulation problem (i.e., the prediction error accumulates over time since the current-frame outputs are used as inputs to the next frame in the auto-regressive fashion) and thus can generate arbitrary long motion dance sequences without freezing effects [Petrovich et al. 2021; Tseng et al. 2023]. The input of our network at each diffusion step $m$ is the noisy group sequence $x_m = \{x_{m,1}^1, ..., x_{m,T}^1; ...; x_{m,1}^N, ..., x_{m,T}^N\}$, however, we skip the $m$ index for ease of notation from now on.

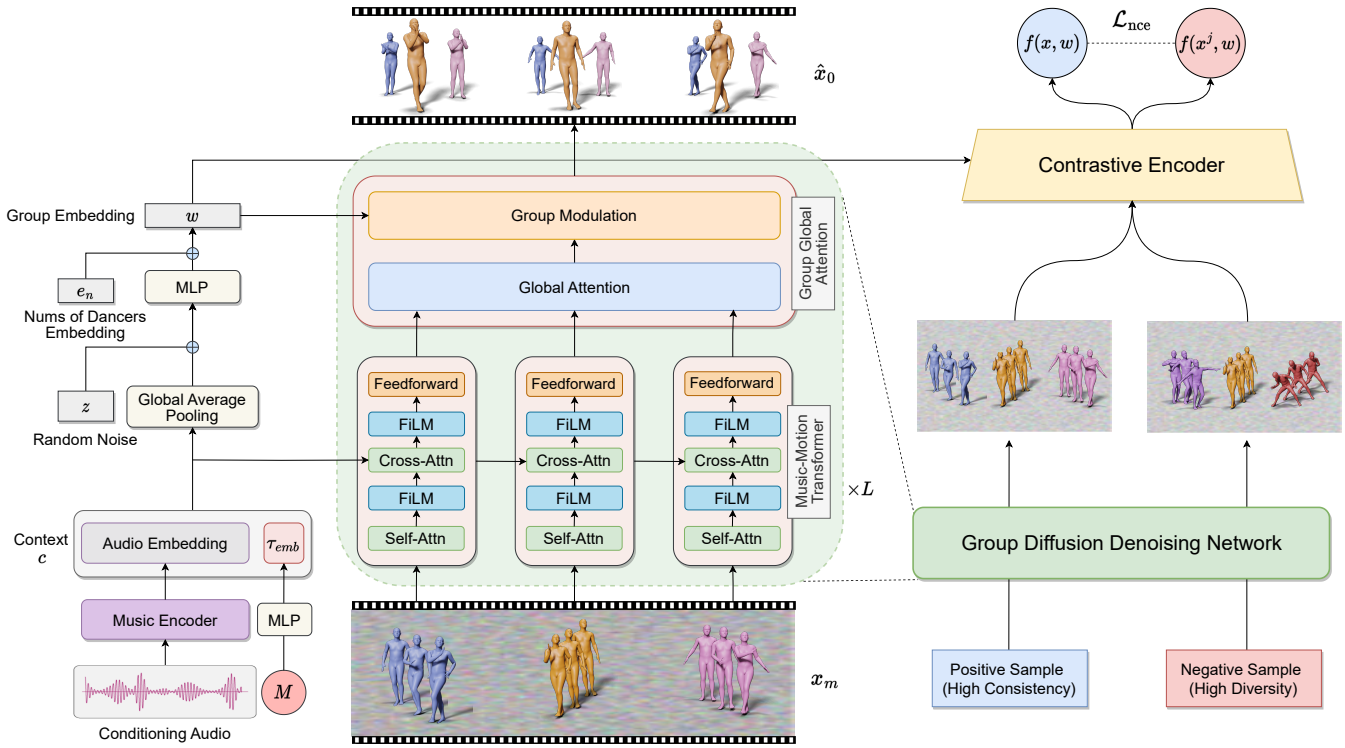


Fig. 2. Detailed illustration of our method for group choreography generation. We adopt a transformer architecture to generate the entire sequence all at once. The input of the denoising network is a noisy group motion sample at each step $m$, along with the conditioning music. The model predicts noiseless sample $\hat{x}_0$, which is then diffused back to $x_{m-1}$ to continue the process until reaching $m = 0$. We further propose to learn the consistency and diversity of samples through a contrastive learning objective with the Contrastive Encoder. The learned encoder is used as guidance signals to control the generation process.

*3.2.1 Music-Motion Transformer.* Given an input extracted audio sequence $a = \{a_1, a_2, ..., a_T\}$, we employ a transformer encoder architecture [Vaswani et al. 2017] to encode the music into the sequence of hidden audio representation $\{c_1, c_2, ..., c_T\}$, which will be used as the conditioning context to the diffusion denoising network. Specifically, we follow the encoder layer as in [Vaswani et al. 2017] which consists of multi-head self-attention layers and feed-forward layers to effectively encode the multi-scale rhythmic patterns and long-term dependencies between music frames. The diffusion time-step $m$ is also projected to the transformer dimension through a separate Multi-layer Perceptron (MLP) with 3 hidden layers to get the embedding $\tau_{emb}$, then concatenated with the music feature sequence to obtain the final conditioning context $c = \{c_1, c_2, ..., c_T, \tau_{emb}\}$.

Although group choreography incorporates the problem of learning the interaction between dancers, we still need to learn the correlation between the dance movements and the accompanying music audio for each dancer. Therefore, we design the Music-Motion Transformer to essentially focus on learning the direct connection between the motion and the music of *each individual dancer* (and not considering the interconnection among dancers yet). Each frame of the noised input motion $x_t^i$ is projected into the transformer dimension by a linear layer followed by an additive positional encoding [Vaswani et al. 2017]. Given the whole group sequence including all dancers $\{x_1^1, ..., x_T^1; ...; x_1^n, ..., x_T^n\}$, we separately encode the motion features of each individual dancer by utilizing the multi-head self-attention [Vaswani et al. 2017] with masking strategy. We implement the masked self-attention (MSA) mechanism as follows:

$$\text{MSA}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + m_{local}\right)V, \tag{6}$$

$$Q = xW^Q, \quad K = xW^K, \quad V = xW^V \tag{7}$$

where $W^Q, W^K \in \mathbb{R}^{d \times d_k}$ and $W^V \in \mathbb{R}^{d \times d_v}$ are learnable projection matrices to transform the input to query, key, and value, respectively. $m_{local}$ is the local attention mask illustrated in Figure 3a. This mask ensures each individual can only attend to their own motion. Subsequently, to incorporate the music conditioning context $c$ into each individual motion features, we adopt a transformer decoder architecture [Vaswani et al. 2017] with cross-attention mechanism (CA) [Saharia et al. 2022; Tseng et al. 2023; Vaswani et al. 2017], where the motion is the query and the music is the key/value.

$$\text{CA}(\tilde{Q}, \tilde{K}, \tilde{V}) = \text{softmax}\left(\frac{\tilde{Q}\tilde{K}^\top}{\sqrt{d_k}}\right)\tilde{V}, \tag{8}$$

$$\tilde{Q} = \tilde{x}\tilde{W}^Q, \quad \tilde{K} = c\tilde{W}^K, \quad \tilde{V} = c\tilde{W}^V \tag{9}$$

where $\tilde{x}$ is the output activation of the MSA block, and $\tilde{W}^Q$, $\tilde{W}^K$, $\tilde{W}^V$ are the learnable projection matrices that have similar behavior to the MSA mechanism.

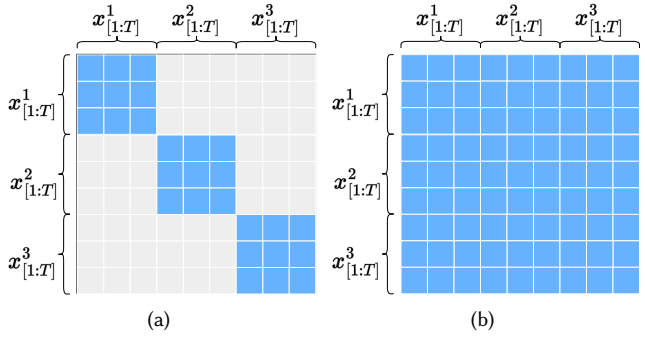


Fig. 3. The local attention mask $m_{local}$ (a) and global attention mask $m_{global}$ (b). The blue cell indicates where frames can attend to each other. Blue color represents zero value of the mask while gray color represents minus infinity. $x^i_{[1:T]}$ indicate motion sequence of $i$-th dancer.

*3.2.2 Group Global Attention.* To ensure the coherency and non-collision in the movements of *all dancers* within the group, such that their dances should correlate with each other under the music condition instead of dancing asynchronously , we first perform global attention via a masked attention mechanism similar to Equation 6 with a full masking strategy $m_{global}$. The attention mask is illustrated in Figure 3b. It allows a dancer to fully attend to all other dancers under the global receptive field. Then, we propose the Group Modulation to enforce the group constraints within the group embedding information.

Inspired by StyleGan [Karras et al. 2019], in which the synthesized image can be manipulated via a latent style vector, we aim to learn a group embedding information from the input music in order to control the group dance generation process. We first apply temporal average pooling to the encoded music feature sequence to obtain a compact representation of the input music $\bar{c} = \frac{1}{T}\sum_{t=1}^{T} c_t$. To increase the variation and diversity of the group information (i.e., avoid limiting the group embedding to only one style of the input music), we inject a random noise drawn from a standard gaussian distribution $z \sim \mathcal{N}(0, I)$ into $\bar{c}$. We use an 8-layer MLP to learn a mapping from the audio representation to the group embedding. We also add a learnable embedding token $e_n$ from a variable-size lookup table $E \in \mathbb{R}^{N\times D}$ up to $N$ maximum dancers, to represent the variation of dancers in the sequence since each sequence may contain different number of dancers. In summary, the process can be written as follows:

$$w = \text{MLP}\left(z + \frac{1}{T}\sum_{t=1}^{T} c_t\right) + e_n, \quad z \sim \mathcal{N}(0, I) \tag{10}$$

**Group Modulation.** To better apply the group information constraints to the learned hidden features of the dancers, we adopt a Group Modulation layer that learns to adaptively influence the output of the transformer attention block by applying an affine transformation to the intermediate features based on the group embedding $w$. More specifically, we utilize two separate linear layers to learn the affine transformation parameters $\{S(w); b(w)\} \in \mathbb{R}^d$ from the group embedding $w$. The predicted affine parameters are then used to modulate the activations sequence $h = \{h_1^1 \dots h_T^1; \dots; h_1^N \dots h_T^N\}$ as follows:

$$\tilde{h} = S(w) * \frac{h - \mu(h)}{\sigma(h)} + b(w) \tag{11}$$

where each channel of the whole activation sequence is first normalized separately by calculating the mean $\mu$ and $\sigma$, and then scaled and biased using the outputted affine parameters $S(w)$ and $b(w)$. Intuitively, this operation shifts the activated hidden motion features of each individual motion towards a unified group representation to further encourage the association between them. Finally, the output features are then projected back to the original motion dimensions via a linear layer, to obtain the predicted outputs $\hat{x}_0$.

## 3.3 Contrastive Diffusion for Controllable Group Dance

*3.3.1 Contrastive Diffusion.* We follow [Oord et al. 2018; Zhu et al. 2023] to learn the representations that encode the underlying shared information between the group embedding information $w$ and the group sequence $x$. Specifically, we model a density ratio that preserves the mutual information between $x$ and $w$ as:

$$f(x, w) \propto \frac{p(x|w)}{p(x)} \tag{12}$$

$f(\cdot)$ is a model (i.e., a neural network) to predict a positive score (how well $x$ is related to $w$) for a pair of $(x, w)$.

To enhance the association between the generated group dance (data) and the group embedding (context), we aim to maximize their mutual information with a Contrastive Encoder $f(\hat{x}, w)$ via the contrastive learning objective as in Equation 13. The encoder takes both the generated group dance sequence $\hat{x}$ and a group embedding $w$ as inputs, and it outputs a score indicating the correspondence between these two.

$$\mathcal{L}_{\text{nce}} = -\mathbb{E}\left[\log \frac{f(\hat{x}, w)}{f(\hat{x}, w) + \Sigma_{x^j \in X'} f(\hat{x}^j, w)}\right] \tag{13}$$

where $X'$ is a set of randomly constructed negative sequences. In general, this loss is similar to the cross-entropy loss for classifying the positive sample, and optimizing it leads to the maximization of the mutual information between the learned context representation and the data [Oord et al. 2018]. Using the contrastive objective, we expect the Contrastive Encoder to learn to distinguish between the two quantities: *consistency* (the positive sequence) and *diversity* (the negative sequence). This is the key factor that enables the ability to control diversity and consistency in our framework.

Here, we will describe our strategy to construct contrastive samples to achieve our target. Recall that we use reverse distribution $p_\theta(x_{t-1}|x_t)$ of Gaussian Diffusion with the mean as the prediction of the model while the variance is fixed to a scheduler (Equation 5). To obtain the contrastive samples, given the true pair is $(x_0, w)$, we first leverage forward diffusion process $q(x_m|x_0)$ to obtain the noised sample $x_m$. Then, our *positive sample* is $p_\theta(x_{m-1}|x_m, w)$. Subsequently, we construct the negative sample from the positive pair by randomly replacing dancers from other group dance sequences $(x_0^j \neq x_0)$ with some probabilities, feeding it through the forward process to obtain $x_m^j$, then our *negative sample* is $p_\theta(x_{m-1}^j|x_m^j, w)$. By constructing contrastive samples this way, the positive pair $(x_0, w)$ represents a group sequence with high consistency, whereas

the negative one represents a high diversity sample. This is because mixing a sample with dancers from different groups is likely to result in substantially distinctive movements between each dancer, making it a group dance sample with high degree of diversity. Note that negative sequences should also match the music because they are motions generated by the network whose inputs are manipulated to increase diversity. Particularly, negative samples are acquired from outputs of the denoising network whose inputs are both the current music and the noised mixed group with some replaced dancers. As the network is trained to reconstruct only positive samples, its outputs will likely follow the music. Therefore, negative samples are not just random bad samples but are the valid group dance generated from the network that is trained to generate group dance conditioned on the music. This is because our main diffusion training objective (Section 4.1.2) is calculated only for ground-truth dances (positive samples) that are consistent with the music. Our proposed strategy also allows us to learn a more powerful group representation as it directly affects the reverse process, which is beneficial to maintaining consistency in long-term synthesis.

*3.3.2 Diversity vs. Consistency.* Using the Contrastive Encoder $f(x_m, w)$, we extend the classifier guidance [Dhariwal and Nichol 2021] to control the generation process. Accordingly, we incorporate $f(x_m, w)$ in the contrastive framework to replace the guiding classifier in the original formula, since it provides a score of how consistent the sample is with the group information. In particular, we shift the mean of the reverse diffusion process with the log gradient of the Contrastive Encoder with respect to the generated data as follows:

$$\hat{\mu}_\theta(x_m, m) = \mu_\theta(x_m, m) + \gamma \cdot \Sigma_\theta(x_m, m)\nabla_{x_m} \log f(x_m, w) \quad (14)$$

where $\gamma$ is the control parameter that uses the encoder to enforce consistency and connection with the group embedding. Since the Contrastive Encoder is trained to classify between high-consistency and high-diversity samples, its gradients yield meaningful guidance signals to control the trade-off process. Intuitively, a positive value of $\gamma$ encourages more consistency between dancers while a negative value (which corresponds to shifting the distribution with a negative gradient step) boosts the diversity between each individual dancer.

## 4 EXPERIMENTS

### 4.1 Implementation Details

*4.1.1 Network Parameters.* The hidden layer of all the MLPs consists of 512 units followed by GELU activation. The hidden dimension of all attention layers is set to $d = 512$, and the attention adopts a multi-head scheme [Vaswani et al. 2017] with 8 attention heads. We also use a feature-wise linear modulation (FiLM) [Perez et al. 2018; Tseng et al. 2023] after each attention layer to strengthen the influence of the conditioning context. At the end of each attention block, we append a 2-layer feed-forward network [Vaswani et al. 2017] with a feed-forward size of 1024 to enhance the expressivity of the learned features. We extract the features from the raw audio signal by leveraging the representations from the frozen Jukebox [Dhariwal et al. 2020], a pre-trained generative model for music, to enhance the model's generalization ability to several kinds of in-the-wild music. In total, the Group Diffusion Denoising Network is comprised of $L = 5$ stacked Music-Motion Transformer and Group Global Attention blocks, along with 2 transformer encoder layers to encode the music features. We implement the architecture of the Contrastive Encoder similarly to the Denoising Network but without cross attention since it does not take the music as input. The output sequence of the Contrastive Encoder is then averaged out and fed into an output layer with one unit. We also make the Contrastive Encoder aware of the current step in the diffusion chain by appending the diffusion timestep embedding to the motion sequence so that it can provide correct guidance signals in the sampling process. Overall, our model has approximately $62M$ trainable parameters.

*4.1.2 Training.* To train the denoising diffusion network, we use the "simple" objective as introduced in [Ho et al. 2020].

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0 \sim q(x_0|c), m \sim [1,M]} \left[ \|x_0 - \mathcal{G}_\theta(x_m, m, c)\|_2^2 \right] \quad (15)$$

To improve the physical plausibility and prevent artifacts of the generated motion, we also utilize auxiliary geometric losses similar to [Tevet et al. 2023].

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{vel}\mathcal{L}_{\text{vel}} + \lambda_{\text{foot}}\mathcal{L}_{\text{foot}} \quad (16)$$

In particular, geometric losses mainly consist of *(i)* a joint position loss $\mathcal{L}_{\text{pos}}$ to better constrain the global joint hierarchy via forward kinematics; *(ii)* a velocity loss $\mathcal{L}_{\text{vel}}$ to increase the smoothness and naturalness of the motion by penalizing the difference between the differences between the velocities of the ground-truth and predicted motions; and *(iii)* a foot contact loss $\mathcal{L}_{\text{foot}}$ to mitigate foot skating artifacts and improve the realism of the generated motions by ensuring the feet to stay stationary when ground contact occurs.

Our total training objective is the combination of the "simple" diffusion objective, the auxiliary geometric losses, and the contrastive loss (Equation 13):

$$\mathcal{L} = \mathcal{L}_{\text{simple}} + \mathcal{L}_{\text{geo}} + \lambda_{\text{nce}}\mathcal{L}_{\text{nce}} \quad (17)$$

We train our model on 4 NVIDIA V100 GPUs using Adam optimizer [Kingma and Ba 2014] with a learning rate of 1e−4 and a batch size of 64 per GPU, which took about 7 days for 500k iterations. The models are trained with $M = 1000$ diffusion noising steps and a cosine noise schedule [Nichol and Dhariwal 2021]. During training, group dance motions are randomly sampled with sequence length $T = 150$ at 30 Hz, which corresponds to 5-second pieces of music. For geometric losses, the loss weights are empirically set to $\lambda_{\text{pos}} = 1.0$, $\lambda_{\text{smooth}} = 1.0$, and $\lambda_{\text{foot}} = 0.005$, respectively. For the contrastive loss $\mathcal{L}_{\text{nce}}$, its weight is $\lambda_{\text{nce}} = 0.001$, the probability of replacing dancers for negative sequences is 0.5, and the number of negative samples empirically is selected to 10.

*4.1.3 Testing.* At test time, we use the DDIM sampling technique [Song et al. 2021] with 50 steps to accelerate the sampling speed of the reverse diffusion process. Accordingly, our model can achieve real-time generation at 30 Hz on a single RTX 2080Ti GPU (excluding the music features extracting step), thanks to the parallelization of the Transformer architecture.

To enable long-term generation, we adopt a strategy that is similar to the one described in [Tseng et al. 2023]. Specifically, we divide the input music sequence into multiple overlapping chunks, with

each chunk having a maximum window size of 5 seconds and overlapped by half with the adjacent chunk. The group dance motions are then generated for each chunk along with the corresponding audio. Subsequently, we merge the outputs by blending the overlapped region between two consecutive chunks using spherical linear interpolation, with the interpolation weight gradually decaying from the current chunk to the next chunk. However, for group choreography synthesis, our model generates dance motions for each dancer in random order. Therefore, we need to establish correspondences between dancers across the chunks (i.e., identifying which one of the $N$ dancers in the next chunk corresponds to a dancer in the current chunk). To accomplish this, we organize all dancers in the current chunk into one set and the dancers in the next chunk into another set, forming a bipartite graph between the two chunks. We can then utilize the Hungarian algorithm [Kuhn 1955] to find the optimal matching, where the Euclidean distance between the two pose sequences serves as the matching weights. Our blending technique is applied at each step of the diffusion sampling process, starting from pure noise, thus it allows the model to gradually denoise the chunks to make them compatible for blending.

## 4.2 Experimental Settings

*4.2.1 Dataset.* We use AIOZ-GDance dataset [Le et al. 2023] in our experiments. AIOZ-GDance is a large-scale group dance dataset including paired music and 3D group motions captured from in-the-wild videos using a semi-automatic method, covering 7 dance styles and 16 music genres. We follow the training and testing split as in [Le et al. 2023] in our experiments.

*4.2.2 Evaluation Protocol.* We use the following metrics to evaluate the quality of single dancing motion: Frechet Inception Distance (FID) [Heusel et al. 2017; Li et al. 2021b], Motion-Music Consistency (MMC) [Li et al. 2021b], Generation Diversity (GenDiv) [Huang et al. 2020; Lee et al. 2019; Li et al. 2021b] , Physical Foot Contact score (PFC) [Tseng et al. 2023]. Concretely, FID score measures the realism of individual dance movements against the ground-truth dance. The MMC evaluates the matching similarity between the motion and the music beats, i.e., how well generated dances follow the beat of the music. The generation diversity (GenDiv) is evaluated as the average pairwise distance of the kinetic features of the motions [Onuma et al. 2008]. The PFC evaluates the physical plausibility of the foot movements by calculating the agreement between the acceleration of the character's center of mass and the foot's velocity.

To evaluate the group dance quality, we follow three metrics introduced in [Le et al. 2023]: Group Motion Realism (GMR), Group Motion Correlation (GMC), and Trajectory Intersection Frequency (TIF). In general, the GMR measures the realism between generated and ground-truth group motions by calculating Frechet Inception Distance on the extract group motion features. The GMC evaluates the synchrony between dancers within the generated group by calculating their cross-correlation. The TIF measures how often the generated dancers collide with each other in their dance movements.

*4.2.3 Baselines.* We compare our GCD method with several recent approaches on music-driven dance generation: FACT [Li et al. 2021b], Transflower [Perez et al. 2021], and EDGE [Tseng et al. 2023], all of which are adapted for benchmarking in the context of group dance generation [Le et al. 2023] since the original methods were specifically designed for single-dance. We also evaluate against GDanceR [Le et al. 2023], a recent model specifically designed for generating group choreography.

Table 1. Performance comparison. High Consistency: parameter $\gamma = 1$; High Diversity: parameter $\gamma = -1$; Neutral: parameter $\gamma = 0$

| Method | | FID↓ | MMC↑ | GenDiv↑ | PFC↓ | GMR↓ | GMC↑ | TIF↓ |
|---|---|---|---|---|---|---|---|---|
| FACT [Li et al. 2021b] | | 56.20 | 0.222 | 8.64 | 3.52 | 101.52 | 62.68 | 0.321 |
| Transflower [Perez et al. 2021] | | 37.73 | 0.217 | 8.74 | 3.07 | 81.17 | 60.78 | 0.332 |
| EDGE [Tseng et al. 2023] | | 31.40 | 0.264 | 9.57 | 2.63 | 63.35 | 61.72 | 0.356 |
| GDANCER [Le et al. 2023] | | 43.90 | 0.250 | 9.23 | 3.05 | 51.27 | 79.01 | 0.217 |
| **GCD (Ours)** | High Consistency | 31.48 | **0.272** | 8.78 | 2.55 | 39.22 | **82.01** | **0.115** |
| | Neutral | **31.16** | 0.261 | 10.87 | **2.53** | **31.47** | 80.97 | 0.167 |
| | High Diversity | 33.37 | 0.255 | **11.34** | 2.58 | 35.63 | 78.19 | 0.209 |

## 4.3 Experimental Results

*4.3.1 Quality Comparison.* Table 1 shows a comparison among the baselines FACT [Li et al. 2021b], Transflower [Perez et al. 2021], EDGE [Tseng et al. 2023], GDanceR [Le et al. 2023], and our proposed GCD. The results clearly demonstrate that our default model setting with "neutral" mode outperforms the baselines significantly across all evaluations. We also observe that EDGE, a recent diffusion dance generation model, can yield very competitive performance on single-dance metrics (FID, MMC, GenDiv, and PFC). This suggests the advantages of diffusion approaches in motion generation tasks. However, it is still inferior to our model under several group dance metrics, showing the limitations of single dance methods in the context of group dance creation. Experimental results highlight the effectiveness of our approach in generating high-quality group dance motions.

To complement the quantitative analysis, we present qualitative examples from FACT, GDanceR, and our GCD method in Figure 4. Notably, FACT struggles to deal with the intersection problem, which is reasonable given that it was not originally designed for group dance generation. As a result, the generated motions from FACT lack coordination and synchronization in most cases. While GDanceR shows improvements in terms of motion quality compared to FACT, the generated motions appear floating, unnatural, and sometimes unsynchronized in many cases. These drawbacks indicate that GDanceR's effort on generating group choreography would still require more refinement to produce consistent and cohesive movements among the dancers. In contrast, our method excels in both controlling consistency and promoting diversity among the generated group dance motions. The outputs from our method demonstrate well-coordinated and realistic movements, implying that it can resolve the challenges of maintaining group coherence while delivering visually appealing results more effectively.

Overall, the conducted quantitative analysis and visual comparisons reaffirm the superior performance of our proposed GCD to generate high-quality, synchronized, and visually pleasing group dance motions.


Fig. 4. Comparison between different dance generation methods when generating dancing in groups.

*4.3.2 Diversity and Consistency Analysis.* Table 1 also presents an in-depth analysis of our method's performance across seven evaluation metrics by adjusting the parameter $\gamma$ in Equation 14 to control the consistency and diversity of the generated group choreographies. The findings reveal that our GCD with high consistency setting ($\gamma = 1$), performs better than other settings in terms of MMC, GMC, and TIR metrics, whereas the high diversity setting ($\gamma = -1$) achieves better results in the GenDiv metric. Meanwhile, the default model shows the best performance in both realism metrics (FID and GMR). It can also be seen that the model is relatively robust to the physical plausibility score (PFC) as there are no noticeable differences among the metric in the three settings. This implies that our model is able to create group animation with different consistency or diversity levels without compromising the plausibility of the movements too much. More interestingly, we found that there are indeed positive correlations between the two measures MMC, GMR, and the trade-off parameter. It is clear that these metrics are better when the consistency level increases. This is reasonable as we expect higher correspondence between the motion and the music (MMC) or higher correlation of the group motions (GMR) when the consistency level grows, which also agrees with the definition of these metrics.

Consistency setups in GCD lead to more similar movements between dancers. As a result, this similarity contributes to high scores in MMC, GMR, and TIR metrics. In contrast, the diversity setups can synthesize more complex motions with greater variation between dancers as measured by the GenDiv metric, but this also makes it more challenging to reach high values of FID, MMC, and TIR, compared with other setups. In addition, Figure 6 shows an example of correlations between motion beats and music beats under high consistency and diversity settings. The music beats are extracted using the beat tracking algorithm from the Librosa library [McFee et al. 2015]. Notably, the velocity curves in high consistency setting display relatively similar shapes, whereas in the high diversity scenario, the curves are clearly distinguished among dancers. Despite the greater variations in high diversity setting, we can observe that the generated motions are matched with the music as the music beats are mostly located near the extrema of the motion curves in both settings. The experiment indicates that our model can faithfully capture different aspects of group choreography with different settings, including diversity and synchrony of the motions. This demonstrates the potential of our method towards various dance applications such as dance training or composing. Furthermore, our method can also produce distinctively different animation sequences under the same setting while adhering to the input music. It is also important to note that all three setups of GCD significantly outperform other baseline models. This verifies the effectiveness of our proposed approach and shows that it can create high-fidelity group dance animations in any setting. For a more detailed visualization of the results, please refer to Figure 5 and our accompanying supplementary video.

*4.3.3 Number of Dancers Analysis.* Table 2 provides insights into results obtained when generating arbitrarily different numbers of dancers using our proposed GCD in the neutral setting. In general, FID, GMR, and GMC metrics do not exhibit a clearly strong correlation with the number of generated dancers but display diverse and varied results. The MMC metric consistently shows its stability across all setups.

As the number of generated dancers increases, the generation diversity (GenDiv) decreases while the trajectory intersection frequency (TIF) increases. However, it is worth noting that the differences observed in these metrics are relatively minor compared to those produced by GDanceR [Le et al. 2023]. This implies that our method can effectively control consistency and diversity, significantly reducing the chances of collisions between dancers and maintaining the overall quality of generated group dance motions.

For a detailed visualization of results, please refer to Figure 7 and our accompanying supplementary video. These results underscore the robustness and flexibility of our method in generating group dance motions across varying numbers of dancers while ensuring consistency, diversity, and avoiding collisions between performers.


Fig. 5. Consistency and diversity trade-off. High Consistency: parameter $\gamma = 1$; High Diversity: parameter $\gamma = -1$; Neutral: parameter $\gamma = 0$.

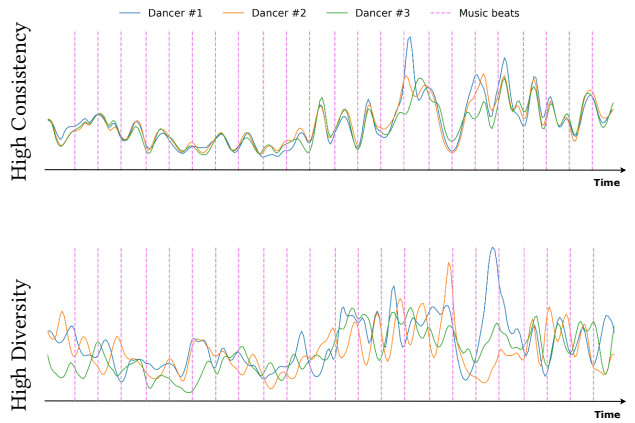


Fig. 6. Correlation between the motion and music beats. The solid curve represents the kinetic velocity of each dancer over time and the vertical dashed line depicts the music beats of the sequence. The motion beats can be detected as the local extrema from the kinetic velocity curve.

*4.3.4 Long-term Analysis.* To evaluate the efficacy of the guidance signal in GCD for creating long-term group dance sequences, we conducted a comparative analysis between GCD and the baseline model GDanceR [Le et al. 2023]. The experiment involved musical pieces of different durations: 15 seconds, 30 seconds and 60 seconds. We show the results with the guidance parameter $\gamma = 0.5$ to enforce consistency with the music over a long duration. For more detailed information, please refer to Figure 9 and our supplementary video.

While both methods produce satisfactory results in the first few seconds of the animations (e.g., about 5-6 seconds), GDanceR starts to exhibit floating and unrealistic movements or freeze into a mean pose in the later period of the sequence. Figure 8 shows the Motion changes comparison between our GCD method and GdanceR. The motion change magnitudes are calculated as average differences of

Table 2. Performance of group dance generation methods when we increase the number of generated dancers, compared with GDanceR. In GCD setup, Neutral mode with $\gamma = 0$ is used.

| Method | #Generated Dancers | FID↓ | MMC↑ | GenDiv↑ | GMR↓ | GMC↑ | TIF↓ |
|---|---|---|---|---|---|---|---|
| GDanceR | 2 | 48.82 | 0.248 | 9.36 | 53.83 | 75.44 | 0.086 |
| | 3 | 44.47 | 0.245 | 9.36 | 55.85 | 74.07 | 0.104 |
| | 4 | 47.32 | 0.248 | 9.24 | 58.79 | 77.71 | 0.162 |
| | 5 | 44.19 | 0.249 | 8.99 | 55.05 | 78.72 | 0.218 |
| **GCD (Ours)** | 2 | 32.62 | 0.266 | 10.41 | 34.09 | 80.26 | 0.067 |
| | 3 | 33.94 | 0.266 | 10.02 | 36.25 | 79.93 | 0.084 |
| | 4 | 35.89 | 0.251 | 9.87 | 36.28 | 81.82 | 0.125 |
| | 5 | 35.08 | 0.264 | 9.92 | 38.43 | 81.44 | 0.168 |


Fig. 7. Group dance generation results of GCD in terms of different numbers of dancers.

the kinetic features [Onuma et al. 2008] between consecutive frames. It is evident that the motion change magnitude of GDanceR is gradually lower and approaching zero in the later half of the 60-second music piece, whereas our method can preserve high magnitudes and variations over time. This is because GDanceR generates almost frozen dance choreographies during this period. In contrast, the group dance motions produced by GCD remain natural with diverse movements throughout the entire duration of all music samples.

These findings confirm that our approach can effectively address the problem of motion generation in long-horizon group dance scenarios. It maintains the motion quality and dynamics of the dance motions, ensuring that the created animations remain visually appealing throughout extended periods. This highlights the advantage of the contrastive strategy to enhance the consistency of the movements of dancers with their group and the music, resulting in significant improvements for long-term dance sequence generation compared to the baseline GDanceR.

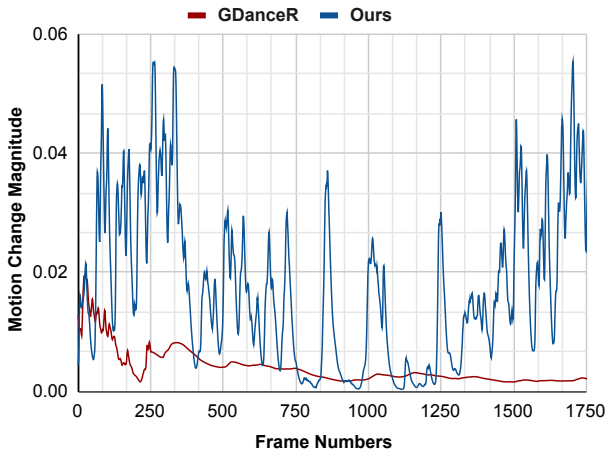


Fig. 8. Motion changes comparison between our proposed GCD and GDanceR. The experiment is conducted on generated group dance results of 60-second music pieces.

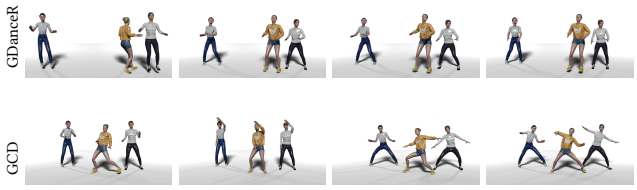


Fig. 9. Long-term results of the 60-second clip. For clearer visualization, please visit our demo video.

#### 4.3.5 *Ablation Analysis.*

**Loss Terms.** The contribution of the geometric loss $\mathcal{L}_{\text{geo}}$ and contrastive loss $\mathcal{L}_{\text{nce}}$ in GCD is thoroughly analyzed and presented in Table 3. The results demonstrate that both losses play a crucial role in enhancing the overall performance across all four evaluation protocols. In particular, it can be seen that the effect of $\mathcal{L}_{\text{geo}}$ on realism metrics (FID, GMR, and PFC) is significant. This observation can be attributed to the fact that this loss improves the physical plausibility and naturalness of the dance motions, empirically mitigating common artifacts such as jittery motion or foot skating. By enforcing the geometric constraints, GCD can generate faithful motions that are on par with real dances. Moreover, the contrastive loss $\mathcal{L}_{\text{nce}}$ contributes positively to the favorable results in the synchrony measures (GMR and GMC). This loss term encourages the model to synchronize the movements of multiple dancers within a group, thus improving the harmonious coordination and cohesion of the generated choreographies. In general, the results of $\mathcal{L}_{\text{geo}}$ and $\mathcal{L}_{\text{nce}}$ validate their importance across various evaluation metrics.

**Group Global Attention.** Results presented in first two lines of Table 3 demonstrate substantial improvements obtained by incorporating Group Global Attention into GCD. It clearly shows that without the Group Global Attention, the performance on the group dance metrics (GMR and GMC) is significantly degraded. We also observe that the removal of this block resulted in inconsistent movements across dancers, where they seem to dance in freestyle without any group choreographic rules and collide with each other in many cases, although they may still follow the rhythm of the music. Results suggest the vital importance of ensuring coherency and regulating collisions for visually appealing group dance animations.

Table 3. Global module contribution and loss analysis. Experiments are conducted on GCD with $\gamma = 0$ (neutral mode).

| Method | FID↓ | MMC↑ | PFC↓ | GMR↓ | GMC↑ |
|---|---|---|---|---|---|
| **GCD** | 31.16 | 0.261 | 2.53 | 31.47 | 80.97 |
| GCD w/o Group Global Attention | 31.35 | 0.263 | 2.62 | 62.23 | 60.72 |
| GCD w/o $\mathcal{L}_{\text{geo}}$ | 39.27 | 0.254 | 2.95 | 37.73 | 80.11 |
| GCD w/o $\mathcal{L}_{\text{nce}}$ | 35.10 | 0.241 | 2.57 | 47.47 | 71.82 |
| GCD w/o ($\mathcal{L}_{\text{geo}}$ & $\mathcal{L}_{\text{nce}}$) | 40.99 | 0.232 | 2.98 | 49.98 | 71.02 |

### 4.4 User Study

Qualitative user studies are important for evaluating generative models as the perception of users tends to be the most relevant metric for many downstream applications. Therefore, we conduct user studies to evaluate our approach in terms of group choreography generation. We organized two separate studies and enlisted roughly 50 individuals with diverse backgrounds to participate in our experiment. Each participant should have some relevant experience in music and dance (at least 1 month of studying or working in dance-related professions). The age of participants varied between 20 and 50, with approximately 55% female and 45% male.

In the initial study, we requested the participants to evaluate the dancing animations based on three criteria: the naturalness of the dancing motions (Realism), how well the movements match the music (Music-Motion Correspondence), and how well the dancers interact or synchronize with each other (Synchronization between Dancers). Participants were asked to rate scores from 0 to 10 for each criterion, ranging from *(0)-very poor*, *(5)-acceptable*, to *(10)-very good*. The collected scores were then normalized to range $[0, 1]$.

This user study encompassed a total of $189 * 3$ samples with songs that are not present in the train set, including those generated from GDanceR [Le et al. 2023], real dance clips from the dataset, and generated results from our proposed method in neutral mode. Figure 10 shows average scores for all mentioned targets across three experiments. Notably, the ratings of our method are significantly higher than GDanceR across all three criteria. We also perform Tukey honest significance tests to determine the significant differences among the three methods. For the first two criteria (Realism and Music-Motion Correspondence), we observe that the mean scores of all methods are significantly different with $p < 0.05$. For synchronization critera, the differences are significant except for the scores between our method and real dances ($p \approx 0.07$). This highlights that our method can even achieve comparable scores with real dances, especially in the synchronization evaluation. This can be attributed to the proposed contrastive diffusion strategy, which can effectively maintain a balance between the consistency of the movements and the group/audio context, as well as diversity in generated dances.

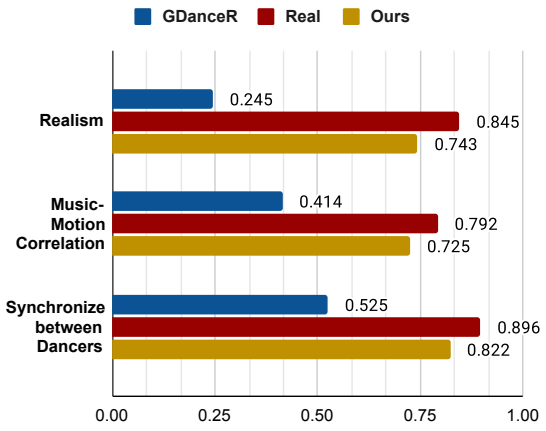


Fig. 10. User study results in three criteria: Realism; Music-Motion Correspondence; and Synchronization between Dancers.

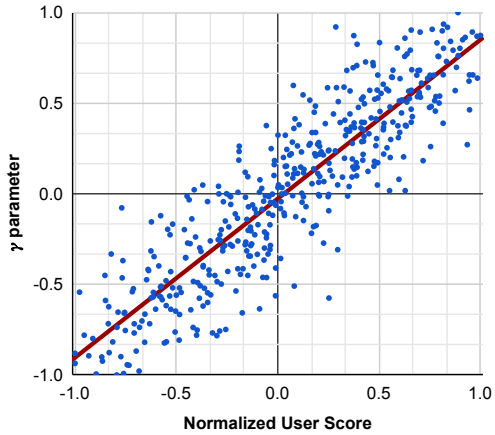


Fig. 11. Correlation between the controlling consistency/diversity and the scores provided by the users .

In the second study, we aim to assess the diversity and consistency of the generated dance outputs and determine if they met the expectations of the users. Specifically, participants were asked to assign scores ranging from 0 to 10 to evaluate the consistency and diversity of each dance clip, i.e., how synchronized or how distinctive movements between dancers does the group dance present. A lower score indicated higher consistency, while a higher score indicated greater diversity. These scores were subsequently normalized to $[-1, 1]$ to align with the studying range of the control parameter employed in our proposed method.

Figure 11 depicts a scatter plot illustrating the relationship between the scores provided by the participants and the $\gamma$ parameter that was used to generate the dance samples. The parameter values were randomly drawn from a uniform distribution with range $[-1, 1]$ to create the animations along with randomly sampled musical pieces. The survey shows a strong correlation between the user scores and the control parameter, in which we calculated the correlation coefficient to be approximately 0.88. The results indicate that the diversity and consistency level of the generated group choreography samples is mostly in agreement with the user evaluation, as indicated by the scores obtained.

## 5 DISCUSSION AND CONCLUSION

While controlling consistency and diversity in group dance generation by using our proposed GCD has numerous advantages and potentials, there are certain limitations. Firstly, it requires tuning the parameters and a complex system that is not trivial to train, to ensure that the generated dance motions can produce the desired level of similarity among dancers while still presenting enough variation to avoid repetitive or monotonous movements. This may involve long inference processes and may require significant computational resources in both the training and testing phases.

Secondly, over-controlling consistency and diversity may introduce constraints on the creative freedom of generated dances. While enforcing consistency can lead to synchronized and harmonious group movements, it may limit the possibility of exploring unconventional or new experimental dance styles. On the other hand, promoting diversity results in unique and innovative dance sequences, but it may sacrifice coherence and coordination among dancers.

Although our model can synthesize semantically faithful group dance animation with effective coordination among dancers, it does not capture clear physical contact between dancers such as hand touching. This is because the data we used in training does not contain such detailed hand motion information. We think that exploring group dance with realistic physical hand interactions is a promising area for future work. Additionally, while our method offers a trade-off between diversity and consistency, achieving perfect alignment between high-diversity movements and music remains a challenging task. The diversity level among dancers and the alignment with music are also heavily influenced by the training data. We believe further efforts are required to reach this.

Lastly, the subjective nature of evaluating consistency and diversity poses a challenge. Metrics for measuring these aspects may not be best fitted. We believe it is essential to consider diverse perspectives and demand domain experts to validate the effectiveness and quality of the generated dance motions.

To conclude, we have introduced GCD, a new method for audio-driven group dance generation that effectively controls the consistency and diversity of generated choreographies. By using contrastive diffusion along with the guidance technique, our approach enables the generation of a flexible number of dancers and long-term group dances without compromising fidelity. Through our experiments, we have demonstrated the capability of GCD to produce visually appealing and synchronized group dance motions. The results of our evaluation, including comparisons with existing methods, highlight the superior performance of our method across various metrics including realism and synchronization. By enabling control over the desired levels of consistency and diversity while preserving fidelity, our work has the potential for applications in entertainment, virtual performances, and artistic expression, advancing the effectiveness of deep learning in generative choreography.

## REFERENCES


Vida Adeli, Ehsan Adeli, Ian Reid, Juan Carlos Niebles, and Hamid Rezatofighi. 2020. Socially and contextually aware human motion and pose forecasting. *IEEE Robotics and Automation Letters (RA-L)* (2020).

Hyemin Ahn, Jaehun Kim, Kihyun Kim, and Songhwai Oh. 2020. Generative autoregressive networks for 3d dancing move synthesis from music. *IEEE Robotics and Automation Letters (RA-L)* (2020).

Alexandre Alahi, Vignesh Ramanathan, and Li Fei-Fei. 2014. Socially-aware large-scale crowd forecasting. In *CVPR*.

Sarah Fdili Alaoui, Cyrille Henry, and Christian Jacquemin. 2014. Physical modelling for interactive installations and the performing arts. *International Journal of Performance Arts and Digital Media* (2014).

Omid Alemi, Jules Françoise, and Philippe Pasquier. 2017. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. *Networks* (2017).

Simon Alexanderson, Rajmund Nagy, Jonas Beskow, and Gustav Eje Henter. 2023. Listen, denoise, action! audio-driven motion synthesis with diffusion models. *ACM Transactions on Graphics (TOG)* (2023).

Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Lars Petersson, and Stephen Gould. 2020. A stochastic conditioning scheme for diverse human motion prediction. In *CVPR*.

Okan Arikan and David A Forsyth. 2002. Interactive motion generation from examples. *ACM Transactions on Graphics (TOG)* (2002).

Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. 2018. Deep motifs and motion signatures. *ACM Transactions on Graphics (TOG)* (2018).

Andreas Aristidou, Anastasios Yiannakidis, Kfir Aberman, Daniel Cohen-Or, Ariel Shamir, and Yiorgos Chrysanthou. 2022. Rhythm is a dancer: Music-driven motion synthesis with global structure. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2022).

Ho Yin Au, Jie Chen, Junkun Jiang, and Yike Guo. 2022. ChoreoGraph: Music-conditioned Automatic Dance Choreography over a Style and Tempo Consistent Dynamic Graph. In *ACM International Conference on Multimedia*.

Daniel Bisig. 2022. Generative Dance-a Taxonomy and Survey. In *International Conference on Movement and Computing*.

Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. 2019. Everybody dance now. In *ICCV*.

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

Wenheng Chen, He Wang, Yi Yuan, Tianjia Shao, and Kun Zhou. 2020. Dynamic future net: Diversified human motion generation. In *ACM International Conference on Multimedia*.

Baptiste Chopin, Hao Tang, and Mohamed Daoudi. 2023. Bipartite Graph Diffusion Model for Human Interaction Generation. *arXiv* (2023).

Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. 2023. MoFusion: A Framework for Denoising-Diffusion-based Motion Synthesis. In *CVPR*.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A generative model for music. *arXiv* (2020).

Prafulla Dhariwal and Alexander Nichol. 2021. Diffusion models beat gans on image synthesis. In *NeurIPS*.

Rukun Fan, Songhua Xu, and Weidong Geng. 2011. Example-based automatic music-driven conventional dance motion synthesis. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2011).

Bin Feng, Tenglong Ao, Zequn Liu, Wei Ju, Libin Liu, and Ming Zhang. 2023. Robust Dancer: Long-term 3D Dance Synthesis Using Unpaired Data. *arXiv* (2023).

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

Bernhard Fink, Bettina Bläsing, Andrea Ravignani, and Todd K. Shackelford. 2021. Evolution and functions of human dance. *Evolution and Human Behavior* (2021).

Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Xinxin Zuo, Zihang Jiang, and Xinchao Wang. 2023. TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration. In *ICCV*.

Wen Guo, Xiaoyu Bie, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. 2022. Multi-person extreme motion prediction. In *CVPR*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *NeurIPS*.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv* (2022).

Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. 2022. AvatarCLIP: zero-shot text-driven generation and animation of 3D avatars. *ACM Transactions on Graphics (TOG)* (2022).

Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, and Mi Zhang. 2020. Dance Revolution: Long Sequence Dance Generation with Music via Curriculum Learning. *CoRR* (2020).

Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. 2023. Diffusion-based Generation, Optimization, and Planning in 3D Scenes. In *CVPR*.

Yin-Fu Huang and Wei-De Liu. 2021. Choreography cGAN: generating dances with music beats using conditional generative adversarial networks. *Neural Computing and Applications* (2021).

Manish Joshi and Sangeeta Chakrabarty. 2021. An extensive review of computational dance automation techniques and applications. *Proceedings of the Royal Society A* (2021).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Pushpajit Khaire and Praveen Kumar. 2022. Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey. *Journal of Visual Communication and Image Representation* (2022).

Sena Kiciroglu, Wei Wang, Mathieu Salzmann, and Pascal Fua. 2022. Long term motion prediction using keyposes. In *3DV*.

Iris Kico, Nikos Grammalidis, Yiannis Christidis, and Fotis Liarokapis. 2018. Digitization and Visualization of Folk Dances in Cultural Heritage: A Review. *Inventions* (2018).

Jihoon Kim, Taehyun Byun, Seungyoun Shin, Jungdam Won, and Sungjoon Choi. 2022a. Conditional motion in-betweening. *Pattern Recognition* (2022).

Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. 2022b. A Brand New Dance Partner: Music-Conditioned Pluralistic Dancing Controlled by Multiple Dance Genres. In *CVPR*.

Jae Woo Kim, Hesham Fouad, and James K Hahn. 2006. Making Them Dance.. In *AAAI Fall Symposium: Aurally Informed Performance*.

Tae-hoon Kim, Sang Il Park, and Sung Yong Shin. 2003. Rhythmic-motion synthesis based on motion-beat analysis. *ACM Transactions on Graphics (TOG)* (2003).

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv* (2014).

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2020. Diffwave: A versatile diffusion model for audio synthesis. *arXiv* (2020).

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. 2002. Motion graphs. In *SIGGRAPH*.

Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* (1955).

Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2023. Music-Driven Group Choreography. In *CVPR*.

Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. In *NeurIPS*.

Lik-Hang Lee, Zijun Lin, Rui Hu, Zhengya Gong, Abhishek Kumar, Tangyao Li, Sijia Li, and Pan Hui. 2021. When Creators Meet the Metaverse: A Survey on Computational Arts. *CoRR* (2021).

Minho Lee, Kyogu Lee, and Jaeheung Park. 2013. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications* (2013).

Buyu Li, Yongchi Zhao, and Lu Sheng. 2022a. DanceNet3D: Music Based Dance Generation with Parametric Motion Transformer. In *AAAI*.

Buyu Li, Yongchi Zhao, Shi Zhelun, and Lu Sheng. 2022b. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jing Li, Di Kang, Wenjie Pei, Xuefei Zhe, Ying Zhang, Zhenyu He, and Linchao Bao. 2021a. Audio2gestures: Generating diverse gestures from speech audio with conditional variational autoencoders. In *ICCV*.

Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021b. Ai choreographer: Music conditioned 3d dance generation with aist++. In *ICCV*.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics (TOG)* (2015).

Shitong Luo and Wei Hu. 2021. Diffusion probabilistic models for 3d point cloud generation. In *CVPR*.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Python in science conference*.

Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. 2018. Single-Shot Multi-person 3D Pose Estimation from Monocular RGB. In *3DV*.

Toan Nguyen, Minh Nhat Vu, Baoru Huang, Tuan Van Vo, Vy Truong, Ngan Le, Thieu Vo, Bac Le, and Anh Nguyen. 2023. Language-Conditioned Affordance-Pose Detection in 3D Point Clouds. *arXiv preprint arXiv:2309.10911* (2023).

Alex Nichol, Heewoo Jun, Prafulla Dhariwal, Pamela Mishkin, and Mark Chen. 2022b. Point-E: A System for Generating 3D Point Clouds from Complex Prompts. *arXiv* (2022).

Alexander Quinn Nichol and Prafulla Dhariwal. 2021. Improved denoising diffusion probabilistic models. In *ICML*.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022a. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *ICML*.

Ferda Ofli, Engin Erzin, Yücel Yemez, and A Murat Tekalp. 2011. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia (TMM)* (2011).

Kensuke Onuma, Christos Faloutsos, and Jessica K Hodgins. 2008. FMDistance: A Fast and Effective Distance Function for Motion Capture Data.. In *Eurographics*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv* (2018).

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *AAAI*.

Guillermo Valle Perez, Jonas Beskow, Gustav Henter, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. 2021. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)* (2021).

Mathis Petrovich, Michael J Black, and Gül Varol. 2021. Action-conditioned 3D human motion synthesis with transformer VAE. In *ICCV*.

Trong-Thang Pham, Nhat Le, Tuong Do, Hung Nguyen, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2023. Style Transfer for 2D Talking Head Animation. *arXiv preprint arXiv:2303.09799* (2023).

Maggi Phillips, Cheryl Stock, and Kim Vincs. 2009. *Dancing between diversity and consistency: Refining assessment in postgraduate degrees in dance.* Western Australian Academy of Performing Arts, Edith Cowan University.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv* (2022).

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *ICML*.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv* (2022).

Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. 2020. Self-supervised dance video synthesis conditioned on music. In *ACMMM*.

Zhiyuan Ren, Zhihong Pan, Xin Zhou, and Le Kang. 2023. Diffusion motion: Generate text-guided 3d human motion by diffusion model. In *ICASSP*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Alla Safonova and Jessica K Hodgins. 2007. Construction and optimal search of interpolated motion graphs. In *SIGGRAPH*.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*.

Junyoung Seo, Wooseok Jang, Min-Seop Kwak, Jaehoon Ko, Hyeonsu Kim, Junho Kim, Jin-Hwa Kim, Jiyoung Lee, and Seungryong Kim. 2023. Let 2D Diffusion Model Know 3D-Consistency for Robust Text-to-3D Generation. *arXiv* (2023).

Yonatan Shafir, Guy Tevet, Roy Kapon, and Amit H Bermano. 2023. Human motion diffusion as a generative prior. *arXiv* (2023).

Nicholas Sharp, Souhaib Attaiki, Keenan Crane, and Maks Ovsjanikov. 2022. Diffusionnet: Discretization agnostic learning on surfaces. *ACM Transactions on Graphics (TOG)* (2022).

Jianxing Shi. 2021. Application of 3D computer aided system in dance creation and learning. In *International Conference on Machine Learning and Big Data Analytics for IoT Security and Privacy*.

Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. 2006. Dancing-to-music character animation. *Computer Graphics Forum* (2006).

Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT with Choreographic Memory. In *CVPR*.

Asako Soga, Bin Umino, and Jeffrey Scott Longstaff. 2005. Automatic composition of ballet sequences using a 3D motion archive. In *1st South-Eastern European Digitization Initiative Conference*.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising Diffusion Implicit Models. In *ICLR*.

Ziyang Song, Dongliang Wang, Nan Jiang, Zhicheng Fang, Chenjing Ding, Weihao Gan, and Wei Wu. 2022. Actformer: A gan transformer framework towards general action-conditioned 3d human motion generation. *arXiv* (2022).

Alexandros Stergiou and Ronald Poppe. 2019. Analyzing human–human interactions: A survey. *Computer Vision and Image Understanding* (2019).

Guofei Sun, Yongkang Wong, Zhiyong Cheng, Mohan S. Kankanhalli, Weidong Geng, and Xiangdong Li. 2020. DeepDance: Music-to-Dance Motion Choreography With Adversarial Learning. *IEEE Transactions on Multimedia (TMM)* (2020).

Jiangxin Sun, Chunyu Wang, Huang Hu, Hanjiang Lai, Zhi Jin, and Jian-Fang Hu. 2022. You Never Stop Dancing: Non-freezing Dance Generation via Bank-constrained Manifold Projection. In *NeurIPS*.

Taoran Tang, Jia Jia, and Hanyang Mao. 2018. Dance with melody: An lstm-autoencoder approach to music-oriented dance synthesis. In *ACMMM*.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. 2023. Human motion diffusion model. In *ICLR*.

Jonathan Tseng, Rodrigo Castellon, and C Karen Liu. 2023. EDGE: Editable Dance Generation From Music. In *CVPR*.

Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. 2019. AIST Dance Video Database: Multi-genre, Multi-dancer, and Multi-camera Database for Dance Information Processing. In *ISMIR*.

Anwaar Ulhaq, Naveed Akhtar, and Ganna Pogrebna. 2022. Efficient Diffusion Models for Vision: A Survey. *arXiv* (2022).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

An Vuong, Minh Nhat Vu, Toan Tien Nguyen, Baoru Huang, Dzung Nguyen, Thieu Vo, and Anh Nguyen. 2023. Language-driven Scene Synthesis using Multi-conditional Diffusion Model. In *NeurIPS*.

Jiashun Wang, Huazhe Xu, Medhini Narasimhan, and Xiaolong Wang. 2021. Multi-Person 3D Motion Prediction with Multi-Range Transformers. In *NeurIPS*.

Zixuan Wang, Jia Jia, Haozhe Wu, Junliang Xing, Jinghe Cai, Fanbo Meng, Guowen Chen, and Yanfeng Wang. 2022. Groupdancer: Music to multi-people dance synthesis with style collaboration. In *ACM International Conference on Multimedia*.

Alexandra Willis, Nathalia Gjersoe, Catriona Havard, Jon Kerridge, and Robert Kukla. 2004. Human movement behaviour in urban spaces: Implications for the design and modelling of effective pedestrian environments. *Environment and Planning B: Planning and Design* (2004).

Jianfeng Xiang, Jiaolong Yang, Binbin Huang, and Xin Tong. 2023. 3D-aware Image Generation using 2D Diffusion Models. *arXiv* (2023).

Nelson Yalta, Shinji Watanabe, Kazuhiro Nakadai, and Tetsuya Ogata. 2019. Weakly-supervised deep recurrent neural networks for basic dance step generation. In *IJCNN*.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2022. Diffusion models: A comprehensive survey of methods and applications. *arXiv* (2022).

Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. 2020. Choreonet: Towards music to dance synthesis with choreographic action unit. In *ACMMM*.

Wenjie Yin, Hang Yin, Kim Baraka, Danica Kragic, and Mårten Björkman. 2022. Dance Style Transfer with Cross-modal Transformer. *arXiv* (2022).

Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. 2022. LION: Latent Point Diffusion Models for 3D Shape Generation. *arXiv* (2022).

Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. 2022. MotionDiffuse: Text-Driven Human Motion Generation with Diffusion Model. *arXiv* (2022).

Li Zhou and Yan Luo. 2022. A Spatio-temporal Learning for Music Conditioned Dance Generation. In *International Conference on Multimodal Interaction*.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019a. On the continuity of rotation representations in neural networks. In *CVPR*.

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. 2019b. Dance dance generation: Motion transfer for internet videos. In *ICCVW*.

Zixiang Zhou and Baoyuan Wang. 2023. Ude: A unified driving engine for human motion generation. In *CVPR*.

Ye Zhu, Kyle Olszewski, Yu Wu, Panos Achlioptas, Menglei Chai, Yan Yan, and Sergey Tulyakov. 2022. Quantized GAN for Complex Music Generation from Dance Videos. In *ECCV*.

Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. 2023. Discrete contrastive diffusion for cross-modal and conditional generation. In *ICLR*.

Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. 2022. Music2Dance: DanceNet for Music-Driven Dance Generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).