Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

An Intelligent Music–Dance Modeling Approach for Assisted Dance Teaching Based on Cycle Consistency

Lun Lu

lulun202602@126.com

Nanjing University

Article

Keywords: Dance teaching, Intelligent dance education, Cycle-consistent learning, Music–motion alignment, Educational motion generation

Posted Date: March 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8621768/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Article

An Intelligent Music–Dance Modeling Approach for Assisted Dance Teaching Based on Cycle Consistency

Lun Lu [1,*]

[1] Nanjing University; lulun202602@126.com
* Correspondence: lulun202602@126.com

Abstract

Dance teaching relies heavily on the close correspondence between music and body movements to help learners develop rhythmic perception, timing control, and expressive skills. However, existing computational approaches for music-driven dance generation often treat music as a simple control input, which limits their ability to capture musical semantics and reduces their effectiveness in educational applications. To address this limitation, this study proposes an **intelligent dance teaching framework based on cycle-consistent music–motion learning**. The proposed framework integrates a forward Music-to-Dance generation model with a reverse Dance-to-Music reconstruction model, forming a cycle-consistent learning process that encourages semantic preservation between music and dance movements. Through this closed-loop mechanism, the generated dance motions are constrained to retain essential musical information, enabling a more meaningful alignment between musical structure and body movement. This property is particularly beneficial in dance teaching scenarios, where understanding the relationship between music and motion is critical for learning. Experimental evaluations conducted on two public dance datasets, AIST++ and PopDanceSet, demonstrate that the proposed approach significantly improves motion quality, rhythmic consistency, and music–dance correspondence compared with existing methods. The results suggest that cycle-consistent music–motion learning provides an effective computational foundation for intelligent dance teaching systems, supporting both instructional demonstration and learner engagement.

Keywords: Dance teaching; Intelligent dance education; Cycle-consistent learning; Music–motion alignment; Educational motion generation

1. Introduction

Dance teaching is a complex educational process that integrates musical perception, bodily coordination, and expressive movement. In traditional dance education, instructors rely heavily on music-driven demonstrations to help learners understand rhythm, timing, and the expressive relationship between music and body movements. Effective dance teaching therefore requires not only accurate movement execution but also a deep alignment between musical structure and physical motion, enabling learners to internalize musical cues through embodied practice.

With the rapid development of artificial intelligence and educational technologies, computational approaches for music-driven dance generation have attracted increasing attention as potential tools for intelligent dance teaching. These methods offer new possibilities for generating instructional demonstrations, supporting repetitive practice, and enhancing learner engagement. However, most existing music-to-dance generation

approaches treat music primarily as a one-way control signal for motion synthesis. While such methods can generate visually plausible movements, they often fail to capture the deeper semantic relationships between music and dance, such as rhythmic structure, musical phrasing, and expressive intent. This limitation reduces their pedagogical value, as learners may imitate surface-level movements without truly understanding the underlying musical–movement correspondence.

From a pedagogical perspective, an effective dance teaching system should ensure that generated movements meaningfully encode musical information and allow learners to perceive how music is translated into motion. In other words, dance movements should not only be synchronized with music but also be capable of conveying musical structure in a way that supports learning. Motivated by this observation, this study introduces a cycle-consistent learning paradigm into music–motion modeling for dance education. By constructing a closed-loop mapping between music and dance, the proposed approach enforces that dance movements generated from music preserve sufficient musical semantics such that the original music can be reconstructed from the movements. This cycle-consistency mechanism serves as a semantic constraint that promotes a deeper and more interpretable alignment between music and motion, which is particularly beneficial in educational contexts.

In this paper, we propose an intelligent dance teaching framework based on cycle-consistent music–motion learning. The framework integrates a forward Music-to-Dance generation model with a reverse Dance-to-Music reconstruction model, forming a closed-loop learning structure that enhances musical understanding in generated dance motions. Experimental evaluations on public dance datasets demonstrate that the proposed approach significantly improves rhythmic consistency, music–motion correspondence, and motion quality compared with existing methods. These results suggest that cycle-consistent music–motion learning provides a promising foundation for intelligent dance teaching systems, supporting both instructional demonstration and learner-centered practice. The remainder of this paper is organized as follows. Section 2 reviews related work in dance teaching and music–motion modeling. Section 3 presents the proposed framework and learning strategy. Section 4 describes the experimental setup and results. Section 5 discusses the educational implications, and Section 6 concludes the paper.

2. Related Works

2.1 AI-Enhanced Dance Teaching and Educational Feedback

Dance teaching is an embodied learning process that requires learners to develop rhythmic perception, temporal coordination, and expressive motor skills through repeated practice and instructor feedback. With the increasing availability of artificial intelligence and sensing technologies, a growing body of research has explored AI-assisted dance education systems to support instruction, practice, and assessment. Early and recent systems primarily rely on pose estimation and motion similarity analysis to provide automated feedback on learner performance. For example, *DancingInside* investigated teacher–AI collaboration in online dance learning environments and highlighted both the potential and limitations of AI-supported instructional feedback through qualitative user studies [1]. Recent reviews further indicate that AI-based feedback in dance education is transitioning from simple scoring mechanisms toward more pedagogically informed feedback design, while still facing challenges in interpretability and instructional alignment [2].

Several application-oriented studies have proposed dance self-learning or remote education systems using joint position and angular similarity to evaluate learner movements and provide corrective feedback [3,4]. Web-based motion capture platforms have also been introduced to support expert-referenced feedback in remote dance education settings, demonstrating improved learner engagement but limited semantic interpretability of feedback signals [5]. In addition, vision-based synchronized movement learning systems, such as SyncUp, have shown that visual computing can support rhythm and

coordination learning, though these systems typically focus on motion synchronization rather than music–movement semantics [6]. Overall, existing AI-assisted dance teaching systems primarily evaluate how closely a learner's movement matches a reference motion, while providing limited insight into how musical structure is embodied in movement.

2.2 Music-to-Dance Generation

Music-driven dance generation has received increasing attention in the multimedia and computer vision communities, driven by advances in deep generative models and the availability of paired music–dance datasets. AIST++ established a widely adopted benchmark for music-conditioned 3D dance generation and introduced the FACT model to capture long-range dependencies between music and motion using cross-modal attention mechanisms [7]. To address long-term stability issues, Dance Revolution employed curriculum learning strategies to improve long-sequence dance generation quality [8].

Subsequent studies have focused on improving motion realism, rhythmic alignment, and controllability. DanceNet modeled music-driven dance generation in an autoregressive manner with explicit rhythm and style conditioning [9], while Transflower adopted probabilistic autoregressive modeling with multimodal attention to increase motion diversity and music correspondence [10]. Bailando further incorporated choreographic memory and actor–critic GPT architectures to enhance long-term coherence and compositional flexibility in dance generation [11]. More recently, multimodal conditioning has been extended to include textual descriptions alongside music, as demonstrated by TM2D, which also introduced new evaluation metrics for motion freezing and alignment consistency [12]. Despite these advances, most music-to-dance generation methods treat music as a one-way conditioning signal and lack mechanisms to explicitly verify whether generated motions encode meaningful musical semantics.

2.3 Diffusion-Based Motion and Dance Generation

Diffusion models have recently emerged as a powerful paradigm for human motion generation, offering improved fidelity and diversity compared to GAN- and VAE-based approaches. MDM adapted classifier-free diffusion models to human motion generation and established a strong baseline for conditional motion synthesis [13]. Subsequent work has explored diffusion-based motion priors and compositional control, enabling more flexible conditioning and longer sequence generation [14,15].

In the context of music-driven dance generation, EDGE combined diffusion models with rich music representations to achieve editable and physically plausible dance motions, accompanied by new evaluation metrics for physical consistency [16]. POPDG introduced the PopDanceSet dataset and demonstrated that diffusion-based modeling improves motion diversity and music–dance alignment for popular dance styles [17]. Listen, Denoise, Action! extended diffusion frameworks to general audio-driven motion synthesis, highlighting the suitability of diffusion models for modeling the inherent one-to-many relationship between audio and motion [18,19]. Hybrid approaches such as MIDGET further combined VQ-VAE and transformer-based architectures to enhance rhythmic matching and motion quality on standard benchmarks [20]. While diffusion models provide strong generative priors, they alone do not guarantee that generated motions preserve high-level musical semantics relevant for instructional use.

Beyond solo dance generation, recent research has begun to address group choreography and multi-dancer coordination, which are common in real-world dance performances and teaching scenarios. AIOZ-GDANCE introduced a large-scale dataset and baseline methods for music-driven group choreography, emphasizing spatial coordination and temporal consistency among multiple dancers [21]. Other studies have focused on refining temporal smoothness and reducing abnormal frames in music-driven dance generation to improve practical usability [22]. These efforts extend the applicability of dance generation models but remain primarily focused on visual coherence and synchronization rather than pedagogical interpretability.

2.4 Dance-to-Music Generation and Bidirectional Modeling

To strengthen cross-modal alignment, an emerging line of research investigates the inverse task of generating music from dance motion. Dance2Music-Diffusion employed latent diffusion models to generate music aligned with dance videos, demonstrating improvements in timbre richness and synchronization [23]. Related work presented at ACM Multimedia explored generating dance music from 3D motion and style labels using latent diffusion frameworks [24]. Earlier Dance2Music studies also provided initial evidence that motion representations can encode musical structure to some extent [25]. Importantly, these bidirectional modeling efforts suggest that if dance motions can reconstruct music, then the motion representations must preserve meaningful musical information. However, most existing studies treat dance-to-music generation as an independent creative task rather than as a semantic constraint for improving music-to-dance learning.

2.5 Cycle Consistency as a Semantic Constraint

Cycle consistency was originally formalized in unpaired image-to-image translation, where enforcing forward–backward reconstruction encouraged models to learn meaningful and information-preserving mappings between domains [26]. This principle has since been widely adopted in cross-modal learning as a form of semantic regularization. In music–motion modeling, cycle consistency provides a natural mechanism to verify whether generated motions retain sufficient musical information beyond shallow cues such as beat alignment. By constraining a music-to-dance model with a reverse dance-to-music reconstruction process, the learning framework can discourage semantically weak solutions and promote deeper alignment between musical structure and embodied movement.

2.6 Research Gap and Our Work

In summary, prior research in AI-assisted dance education has largely focused on pose similarity evaluation and automated feedback, offering limited support for understanding the relationship between musical structure and movement [1–6]. Meanwhile, music-driven dance generation has achieved remarkable progress in motion realism and diversity through autoregressive and diffusion-based models [7–20], yet most approaches rely on one-way conditioning and lack explicit mechanisms to ensure that generated motions encode interpretable musical semantics. Although bidirectional modeling and dance-to-music generation have been explored [23–25], they have not been systematically integrated as a semantic constraint for dance teaching applications.

To bridge this gap, this paper proposes a **cycle-consistent music–motion learning framework for intelligent dance teaching**. By constructing a closed-loop mapping between music and dance and enforcing cycle consistency during training, the proposed approach encourages generated dance motions to preserve reconstructable musical information. This design promotes a deeper and more interpretable alignment between music and movement, which is particularly valuable for instructional demonstration and learner understanding in dance education contexts.

3. Methods

3.1 Overall Framework

The dance motion generation method based on the proposed cycle-consistency framework is illustrated in Figure 1. To clearly present the main network structure, the time steps of the diffusion process are omitted in the figure. The proposed approach first trains a Dance2Music model, and then utilizes the cycle-consistency loss as an auxiliary training objective to train the Music2Dance model, while keeping the parameters of the Dance2Music model fixed during this stage.

The cycle-consistency loss serves as an auxiliary supervision signal to ensure that, after the Music2Dance model generates dance motions, the generated motions—when used as conditional inputs to the Dance2Music model—can produce music that is highly similar to the original input music. The detailed implementation of this process is described in Section 3.3. In addition, considering that aligning representations in the latent space leads to better convergence and stronger generalization ability, this study aligns the

latent representations of the original music and the reconstructed music within the latent space of a variational autoencoder.

The Music2Dance model adopts the multi-scale spatiotemporal hybrid attention-based dance motion generation model. For the Dance2Music model, several mainstream generative model architectures were explored. Experimental results indicate that a combination of a variational autoencoder and a latent diffusion model is the most suitable structure for the current task. The detailed architecture of the Dance2Music model is introduced in Section 3.2, and additional experimental results of the generative models are reported in Section 4.

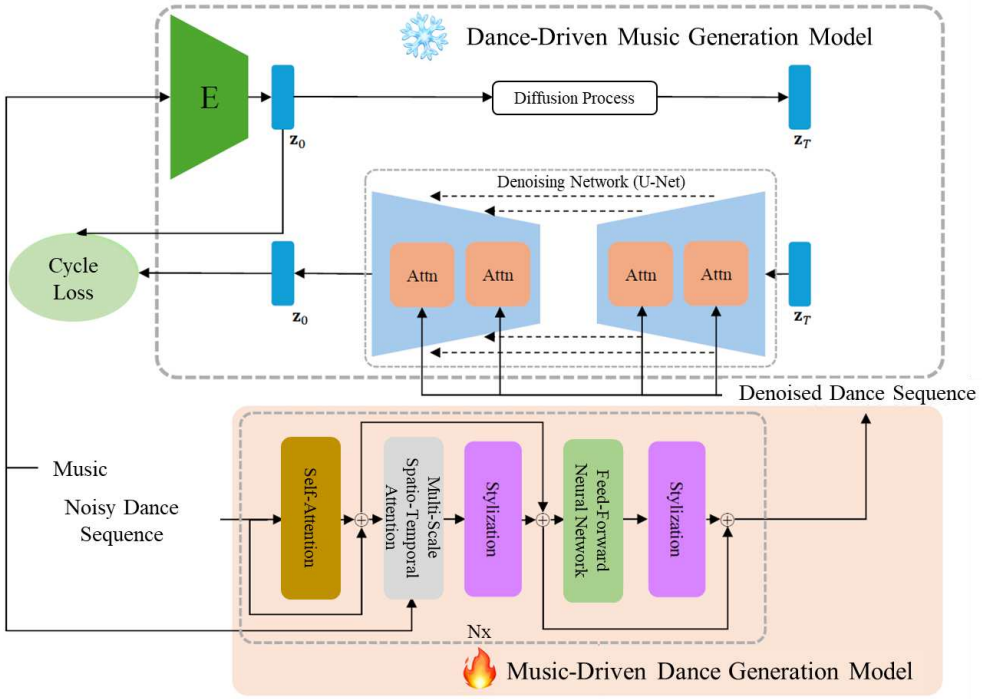


Figure 1. Overall pipeline of the cycle-consistency-based dance motion generation framework.

3.2 Dance-Driven Music Generation Model

The Dance2Music model adopted in this study is illustrated in Figure 2. The model consists of a variational autoencoder (VAE) and a latent diffusion model, which are introduced in detail in Sections 3.2.1 and 3.2.2, respectively. The training process is divided into two stages. In the first stage, the variational autoencoder is trained to learn latent representations of music. In the second stage, the latent diffusion model is trained to learn the data distribution of music latent representations.

Compared with directly modeling the distribution in the original data space, learning the distribution in the latent space is simpler, which reduces the difficulty of optimization and accelerates convergence. Moreover, this latent-space modeling strategy improves the generalization ability of the model, reduces training cost, and ultimately enhances the quality of the generated music.

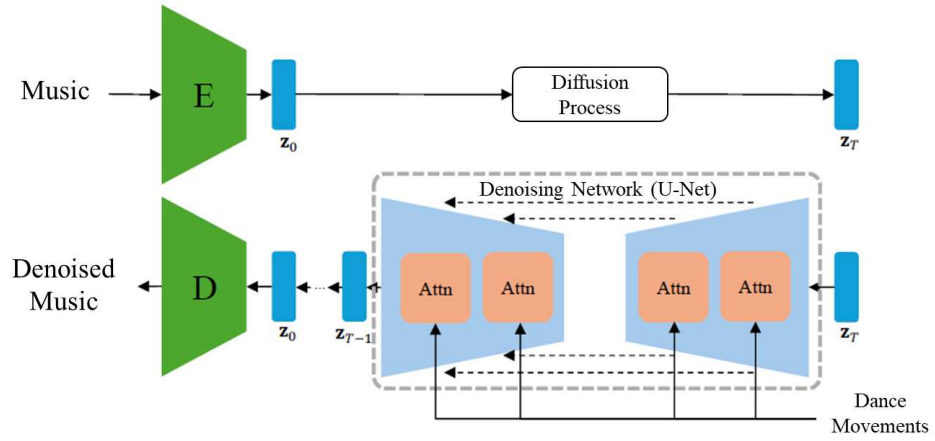


Figure 2 Network Architecture of the Dance-Driven Music Generation Model

3.2.1 Variational Autoencoder

The variational autoencoder (VAE) consists of two components: an **encoder** E and a **decoder** D. The encoder E maps the input data $x \in \mathbb{R}^{F\times T}$ into a probabilistic distribution in the latent space, where F denotes the number of frequency bins and T denotes the temporal dimension. The posterior distribution parameterized by the encoder is denoted as $r_\phi(z \mid x)$, where ϕ represents the parameters of the encoder. Through the encoder, the parameters describing the distribution of the latent variable z are obtained as

$$[\mu, \log\ \sigma^2] = E_\phi(x), \qquad (1)$$

where $\mu \in \mathbb{R}^{d_z}$ and $\log\ \sigma^2 \in \mathbb{R}^{d_z}$ represent the mean vector and the logarithmic variance vector, respectively, and d_z denotes the dimensionality of the latent space. Accordingly, the posterior distribution $r_\phi(z \mid x)$ is defined as

$$r_\phi(z \mid x) = \mathcal{N}(z \mid \mu, \mathrm{diag}(\sigma^2)). \qquad (2)$$

To sample from this distribution while ensuring that gradients can be propagated through the stochastic sampling process, the reparameterization trick is applied:

$$z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I), \qquad (3)$$

where $\odot$ denotes element-wise multiplication (Hadamard product). Given a latent variable z, the decoder D defines the conditional probability distribution $g_\psi(x \mid z)$, where $g_\psi(x \mid z)$ is assumed to follow a Gaussian distribution and ψ represents the parameters of the decoder. The decoder maps the latent variable z back to the original data space as

$$\hat{x} = D_\psi(z), \qquad (4)$$

where $\hat{x} \in \mathbb{R}^{F\times T}$ denotes the reconstructed Mel-spectrogram.

The loss function of the variational autoencoder consists of two components:

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\mathrm{recon}} + \mathcal{L}_{\mathrm{KL}}, \qquad (5)$$

where the reconstruction loss measures the discrepancy between the original input and the reconstructed output:

$$\mathcal{L}_{\mathrm{recon}} = -\mathbb{E}_{r_\phi(z|x)}\left[\log\ g_\psi(x \mid z)\right], \qquad (6)$$

and the Kullback–Leibler (KL) divergence term encourages the posterior distribution generated by the encoder to be close to the prior distribution $\pi(z)$:

$$\mathcal{L}_{\mathrm{KL}} = D_{\mathrm{KL}}\left(r_\phi(z \mid x) \parallel \pi(z)\right). \qquad (7)$$

In this work, the prior distribution is chosen as the standard normal distribution $\pi(z) = \mathcal{N}(0, I)$, for which the KL divergence admits a closed-form solution:

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2}\sum_{i=1}^{d_z}(\mu_i^2 + \sigma_i^2 - \log\ \sigma_i^2 - 1)\,. \qquad (8)$$

Here, d_z denotes the dimensionality of the latent variable z. By minimizing the overall loss function, the variational autoencoder simultaneously learns a compact latent representation of the data and a generative model capable of reconstructing the original input.

3.2.2 Latent Diffusion Model

Although variational autoencoders have limited generative capacity, they provide an effective mechanism for compressing high-dimensional data into a low-dimensional latent space. Leveraging the representational power of the variational autoencoder, the latent diffusion model applies the diffusion process within the latent space learned by the VAE, enabling efficient and high-quality music generation.

Specifically, the latent diffusion model uses a pretrained variational autoencoder to map the high-dimensional input data $x$into a low-dimensional latent representation z_0:

$$\mu, \log \sigma^2 = E_\phi(x), \quad (9)$$

$$z_0 = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I). \quad (10)$$

After obtaining the latent representation z_0, the latent diffusion model progressively adds noise to $z_0$according to a predefined noise schedule parameter β_t:

$$z_t = \sqrt{1-\beta_t}\, z_{t-1} + \sqrt{\beta_t}\, \epsilon, \epsilon \sim \mathcal{N}(0, I), \quad (11)$$

where $t \in [1, T]$denotes the diffusion timestep and $\beta_t$is the noise variance parameter that increases with t. When $t = T$, the latent variable $z_T$approaches pure Gaussian noise.

By unrolling the Markov chain, the latent representation at an arbitrary timestep $t$can be directly obtained from z_0:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \epsilon \sim \mathcal{N}(0, I), \quad (12)$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s = \prod_{s=1}^{t} (1-\beta_s)$.

The model learns the reverse diffusion process to achieve conditional generation, i.e., to recover the latent representation from noisy observations. Specifically, a denoising network $\hat{z}_0$is trained to predict the original latent representation z_0. This network takes the noisy latent variable z_t, the timestep t, and the conditional information $c$as inputs. The diffusion loss is defined as

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{t, z_0, \epsilon, c}[\| \hat{z}_0(z_t, t, c) - z_0 \|^2]. \quad (13)$$

During inference, the sampling process starts from random Gaussian noise $z_T \sim \mathcal{N}(0, I)$and gradually denoises it to obtain the latent representation z_0. In this study, the **DDIM** (Denoising Diffusion Implicit Models) sampling strategy is adopted, which enables high-quality generation with fewer sampling steps:

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{z}_0 + \sqrt{\frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}} \left(z_t - \sqrt{\bar{\alpha}_t}\, \hat{z}_0\right). \quad (14)$$

Here, $\hat{z}_0$denotes the original latent representation predicted by the denoising network.

After obtaining the latent representation z_0, the pretrained VAE decoder $D_\psi$is used to map the generated latent representation back to the original data space:

$$\hat{x} = D_\psi(z_0). \quad (15)$$

3.3 Cycle-Consistency Framework

3.3.1 Framework Architecture

This section presents the dance motion generation method based on the proposed **cycle-consistency framework** in detail. The framework integrates the **music-driven dance generation model (Music2Dance)**, which is based on multi-scale spatiotemporal hybrid attention, with the **dance-driven music generation model (Dance2Music)**. Together, these two models form a complete **music–dance–music** cycle generation process. This framework enables the Music2Dance model to capture deeper semantic information from music during dance generation, facilitating a more comprehensive exploration of the complex relationships between music and dance, and ultimately improving the quality of the generated dance motions.

As illustrated in **Figure 1**, the framework consists of two main components:

- **Music-driven dance generation model (Music2Dance):** This model is based on the multi-scale spatiotemporal hybrid attention mechanism and is capable of generating high-quality dance motions with strong physical plausibility and accurate beat alignment.
- **Dance-driven music generation model (Dance2Music):** This model, described in Section 3.2, is an audio generation framework based on a variational autoencoder

and a latent diffusion model, which generates music conditioned on dance motion sequences.

In addition, two different representations are adopted for music features within the framework:

- The Music2Dance model utilizes high-level music features extracted by the **Jukebox** large-scale music model, denoted as $f_{\text{juke}} \in \mathbb{R}^{T \times D_j}$, where D_j represents the dimensionality of the Jukebox feature.
- The Dance2Music model uses the **Mel-spectrogram** $m \in \mathbb{R}^{F \times T}$ as the music representation, where F denotes the number of Mel frequency bins.

This design leverages the complementary advantages of different music representations: Jukebox features encode richer high-level semantic information, which facilitates deeper semantic understanding in the Music2Dance model, while Mel-spectrograms provide a more direct audio representation suitable for music generation.

3.3.2 Cycle-Consistency Mechanism

Given the Jukebox music features f_{juke} of the original music and the corresponding Mel-spectrogram m, a dance motion sequence $d \in \mathbb{R}^{T \times D}$ is generated through the Music2Dance model:

$$d = \text{Music2Dance}(f_{\text{juke}}), \quad (16)$$

where D denotes the dimensionality of the dance motion features and T denotes the temporal length of the audio sequence.

The generated dance sequence d is then fed into the Dance2Music model to reconstruct the Mel-spectrogram of the music:

$$\hat{m} = \text{Dance2Music}(d). \quad (17)$$

The **cycle-consistency loss** measures the discrepancy between the original music m and the reconstructed music $\hat{m}$ in the latent space of the variational autoencoder:

$$\mathcal{L}_{\text{cycle}} = \| E_\phi(m) - E_\phi(\hat{m}) \|^2. \quad (18)$$

Here, E_ϕ denotes the encoder of the variational autoencoder introduced in Section 3.2.1. Compared with computing differences directly in the Mel-spectrogram space, measuring the discrepancy in the latent space of the variational autoencoder enables better alignment of musical semantic differences while suppressing irrelevant surface-level details.

3.3.3 Training Strategy

The proposed cycle-consistency framework is trained in two stages.

1. **Pretraining stage:**
   The **Dance2Music** model is first trained independently, enabling it to generate corresponding music from a given dance motion sequence.
2. **Cycle-consistency training stage:**
   (a) The parameters of the Dance2Music model are fixed, and only the parameters of the **Music2Dance** model are updated.
   (b) By minimizing the cycle-consistency loss, the Music2Dance model is guided to generate dance motions that preserve sufficient musical information, thereby deepening its understanding of musical semantics and enabling the Dance2Music model to accurately reconstruct the original music.

This two-stage training strategy effectively avoids instability during joint optimization and simultaneously enhances the Music2Dance model's ability to learn deep semantic information from music.

3.3.4 Loss Function

The overall loss function of the Music2Dance model during the cycle-consistency training stage is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{hybrid}} + \lambda_{va}\mathcal{L}_{va} + \lambda_{joint}\mathcal{L}_{joint} + \lambda_{phy}\mathcal{L}_{phy} + \lambda_{cycle}\mathcal{L}_{cycle}. \quad (19)$$

where:

- $\mathcal{L}_{\text{hybrid}}$ is the basic diffusion loss of the Music2Dance model, which guides the model to recover dance motions from noise.

- $\mathcal{L}_{va}$is the velocity–acceleration loss, which ensures that the velocity and acceleration of the generated dance motions are consistent with those of real dance sequences.
- $\mathcal{L}_{joint}$is the joint position loss, which enforces the accuracy of dance motions in three-dimensional space.
- $\mathcal{L}_{phy}$is the physical plausibility loss, which reduces the generation of motions that violate physical constraints.
- $\mathcal{L}_{cycle}$is the cycle-consistency loss, which enforces semantic consistency between the generated dance motions and the original music.

The hyperparameters λ_{va}, λ_{joint}, λ_{phy}, and $\lambda_{cycle}$are used to balance the contributions of the individual loss terms.

4. Results

4.1 Experimental Environment and Datasets

In terms of hardware configuration, four **NVIDIA L40S GPUs** were employed, each equipped with **48 GB** of memory. The large GPU memory enables the processing of large-scale data batches and accelerates the training process.

On the software side, the operating system used was **Ubuntu 24.04 LTS**, which is widely adopted in deep learning research. The experiments were conducted using **Python 3.8**, and the deep learning framework was **PyTorch 1.12**, which has become one of the most commonly used frameworks in the research community. In addition, **CUDA 12.2** was utilized to further accelerate model training.

This study focuses on the task of **music-driven 3D dance motion generation** and employs two publicly available datasets: **AIST++** and **PopDanceSet**.

The **AIST++** dataset contains **10 different dance styles**, with dances performed by **30 dancers**. The training and test sets include **952** and **20** dance sequences, respectively, along with their corresponding music. In terms of motion representation, AIST++ supports both **17-joint COCO-format** motion representations and **24-joint SMPL model parameters** for representing dance motion sequences.

Compared with AIST++, the **PopDanceSet** dataset includes **19 different dance styles**, recorded by **132 dancers**, and features more complex dance motions that are closer to real-world dance scenarios. The training and test sets contain **1012** and **24** dance sequences, respectively, with corresponding music. Similar to AIST++, PopDanceSet supports both **17-joint COCO-format** motion representations and **24-joint SMPL-based** motion representations.

4.2 Evaluation Metrics

The quality of generated music is evaluated from two aspects: **rhythmic alignment** and **content consistency**.

For rhythmic evaluation, the metrics include **Beats Coverage Score (BCS)** , **Beats Hit Score (BHS)** , and **Beats Align Score (BAS)** . Among these, BAS measures the degree of alignment between the beats of the generated music and those of the dance motions used as conditions.

The Beats Coverage Score (BCS) measures the coverage of beats in the generated music relative to the beats in the reference music and is defined as the ratio between the number of detected beats in the generated music and the number of beats in the reference music:

$$\mathrm{BCS} = \frac{N_{\mathrm{syn}}}{N_{\mathrm{ref}}} = \frac{\sum_{i=1}^{n} b_i^{\mathrm{syn}}}{\sum_{i=1}^{n} b_i^{\mathrm{ref}}}, \qquad (20)$$

where $b^{\mathrm{syn}} \in \{0,1\}^n$ denotes the beat sequence of the generated music, $b^{\mathrm{ref}} \in \{0,1\}^n$denotes the beat sequence of the reference music, and $n$is the length of the audio sequence.

The Beats Hit Score (BHS) measures the accuracy of beat alignment between the generated music and the reference music. It is defined as the ratio between the number of

time steps at which beats occur simultaneously in both the generated and reference music and the total number of beats in the reference music:

$$\mathrm{BHS} = \frac{N_{\mathrm{hit}}}{N_{\mathrm{ref}}} = \frac{\sum_{i=1}^{n} \mathbb{I}(b_i^{\mathrm{ref}} = 1 \wedge b_i^{\mathrm{syn}} = 1)}{\sum_{i=1}^{n} b_i^{\mathrm{ref}}}, \quad (21)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function, which takes the value 1 if the condition is satisfied and 0 otherwise. A higher BHS value indicates better temporal alignment between the beats of the generated music and the reference music.

To evaluate content consistency, **Frechet Audio Distance (FAD)** is employed. FAD measures the similarity between generated music and reference music in terms of content and is inspired by the Frechet Inception Distance (FID) used in the image domain. It evaluates content differences by comparing the distributions of generated and reference audio in a learned feature space.

Specifically, a pretrained audio feature extractor **VGGish** is used to extract feature embeddings from the generated and reference audio, resulting in embeddings $X_{\mathrm{gen}} \in \mathbb{R}^{n_1 \times d}$ and $X_{\mathrm{ref}} \in \mathbb{R}^{n_2 \times d}$, where n_1 and n_2 denote the number of samples and d denotes the feature dimensionality.

These feature embeddings are assumed to follow multivariate Gaussian distributions $\mathcal{N}(\mu_{\mathrm{gen}}, \Sigma_{\mathrm{gen}})$ and $\mathcal{N}(\mu_{\mathrm{ref}}, \Sigma_{\mathrm{ref}})$, where:

$$\mu_{\mathrm{gen}} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{\mathrm{gen},i}\,, \Sigma_{\mathrm{gen}} = \frac{1}{n_1}\sum_{i=1}^{n_1}(X_{\mathrm{gen},i} - \mu_{\mathrm{gen}})(X_{\mathrm{gen},i} - \mu_{\mathrm{gen}})^T, \quad (22)$$

$$\mu_{\mathrm{ref}} = \frac{1}{n_2}\sum_{i=1}^{n_2} X_{\mathrm{ref},i}\,, \Sigma_{\mathrm{ref}} = \frac{1}{n_2}\sum_{i=1}^{n_2}(X_{\mathrm{ref},i} - \mu_{\mathrm{ref}})(X_{\mathrm{ref},i} - \mu_{\mathrm{ref}})^T. \quad (23)$$

The Frechet Audio Distance is then computed as:

$$\mathrm{FAD} = \| \mu_{\mathrm{ref}} - \mu_{\mathrm{gen}} \|^2 + \mathrm{Tr}(\Sigma_{\mathrm{ref}} + \Sigma_{\mathrm{gen}} - 2\sqrt{\Sigma_{\mathrm{ref}}\Sigma_{\mathrm{gen}}}), \quad (24)$$

where $\|\cdot\|$ denotes the Euclidean norm and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. A lower FAD value indicates that the generated audio distribution is closer to the reference distribution, implying higher content similarity. When the two distributions are identical, the FAD value equals zero.

4.3 Implementation Details

The implementation details of the **Dance2Music** model are described as follows.

The encoder and decoder of the variational autoencoder are both based on convolutional neural networks. The convolution kernels have a size of 3×3, a stride of 1, and padding of 1. The encoder consists of four convolutional downsampling blocks with channel sizes of 128, 256, 512, and 512, respectively. Downsampling is achieved using convolution layers with a stride of 2 to reduce the spatial resolution of the feature maps. The decoder consists of four convolutional upsampling blocks with channel sizes of 512, 512, 256, and 128, respectively. Upsampling is implemented using interpolation followed by convolution to restore the spatial resolution of the feature maps.

The denoising network of the latent diffusion model adopts a U-shaped convolutional neural network architecture. Cross-attention mechanisms are used to incorporate conditional information, and ResNet blocks are employed for feature extraction. The timestep is represented by a 512-dimensional vector. The four downsampling blocks have channel sizes of 128, 256, 512, and 512, with downsampling achieved via stride-2 convolutions. The four upsampling blocks have channel sizes of 512, 512, 256, and 128, with upsampling implemented via interpolation followed by convolution. In addition, skip connections are introduced between the encoder and decoder to preserve fine-grained details.

For optimization, the **AdamW** optimizer is used with a learning rate of 0.0001, beta values of 0.95 and 0.999, a weight decay coefficient of 1×10^{-6}, and an epsilon value of 1×10^{-8}.

Training is conducted on a single **NVIDIA RTX 4090** GPU with a batch size of 64. During training, the input music Mel-spectrogram is denoted as $m \in \mathbb{R}^{256\times T}$, and the conditional dance motion sequence is denoted as $d \in \mathbb{R}^{T\times 156}$. The first three dimensions of the dance representation correspond to root translation, followed by the 6-degree-of-freedom rotational representations of 24 joints in the SMPL human body model. The remaining dimensions represent binary foot, hand, and head contact signals.

4.4 Comparative Experiments

The comparative experimental results with state-of-the-art methods are presented in **Table 4.1**. The proposed cycle-consistency-based dance motion generation method achieves state-of-the-art performance on both the **PopDanceSet** and **AIST++** datasets.

On the **PopDanceSet** dataset, the proposed method outperforms previous state-of-the-art approaches in terms of **dance quality**, **motion diversity**, and **dance–music alignment**. By modeling the complex relationship between music and dance more effectively, the proposed method improves physical plausibility and motion realism. Specifically, the physical consistency metric **PFC** and the global dance physical consistency metric **PBC** are improved by **14%** and **21%**, respectively, compared with the previous best-performing method. Motion diversity also reaches a level comparable to the strongest baseline. Moreover, the cycle-consistency mechanism plays a crucial role in enhancing dance–music alignment, achieving a **BAS score of 0.248**, which is significantly higher than existing methods.

On the **AIST++** dataset, except for the **Div_vh** metric, the proposed method outperforms previous approaches across all evaluation metrics. Compared with the best baseline, the proposed method improves **PFC**, **PBC**, **Div_vh**, and **BAS** by **0.75%**, **12.51%**, **7.78%**, and **0.42%**, respectively. Compared with rhythm-focused methods such as **EDGE** and **POPDG**, the proposed approach achieves more balanced performance. Although **Div_vh** is slightly lower than the best-performing method by **0.031**, this trade-off results from encouraging smoother and more physically plausible motion patterns, thereby reducing abrupt transitions that violate human biomechanics.

These results demonstrate that the proposed method effectively captures deeper semantic relationships between music and dance, leading to improved dance quality, diversity, and music–dance alignment.

Table 1 Comparative Results with State-of-the-Art Methods

Dataset	Method	Motion Quality		Motion Diversity		Motion–Music Alignment
		PFC ↓	PBC ↓	Div_vc ↑	Div_vh ↑	BAS ↑
PopDanceSet	Ground Truth	1.5824	8.7365	9.0219	7.2931	0.174
	Bailando	3.9751	4.8863	5.1835	5.4342	0.230
	EDGE	3.8366	4.0348	6.1709	5.7568	0.224
	POPDG	1.8253	5.9492	7.1342	5.8314	0.233
	Ours	**1.5666**	**7.2218**	**6.0455**	**6.1340**	**0.248**
AIST++	Ground Truth	1.1806	8.0353	8.5190	8.9384	0.453
	Bailando	1.0633	6.3632	14.8836	6.6855	0.474
	EDGE	1.0792	5.1172	2.5946	2.4477	0.466
	POPDG	0.9952	4.7050	2.4775	2.7832	0.468
	Ours	**0.9877**	**7.1591**	**2.7965**	**2.7441**	**0.476**

4.5 Ablation Studies

To validate the effectiveness of the proposed approach, ablation experiments are conducted on the **PopDanceSet** dataset. The baseline model is the multi-scale spatiotemporal hybrid attention-based dance generation model, while the full model incorporates the proposed cycle-consistency framework.

The ablation results are shown in **Table 2**. Compared with the baseline model, the cycle-consistency-enhanced model achieves consistent improvements in **dance quality**, **motion diversity**, and **dance–music alignment**, with particularly notable gains in alignment. The **BAS score increases from 0.239 to 0.248**, significantly surpassing previous state-of-the-art methods. These results indicate that the cycle-consistency mechanism encourages the generated dance motions to encode richer rhythmic and semantic information from music, leading to more coherent and physically plausible dance sequences.

In addition, to further analyze the effectiveness of different generative backbones in the dance-driven music generation task, this study evaluates several alternative architectures, including **GAN**, **Diffusion**, **VAE-GAN**, and **VAE-Diffusion**. Since PopDanceSet and AIST++ are music-driven dance generation datasets, PopDanceSet is used as the benchmark for evaluating music generation performance.

Table 2 Ablation Results on PopDanceSet

Baseline	Cycle-Consistency	PFC ↓	PBC ↓	Div_vc ↑	Div_vh ↑	BAS ↑
✓	✗	1.7602	6.8442	5.2782	6.1855	0.239
✓	✓	1.5666	7.2218	6.0455	6.1340	0.248

Table 3 Performance of Different Generative Architectures on Dance-Driven Music Generation

Method	BCS ↑	BHS ↑	FAD ↓	BAS ↑
Ground Truth	100.0	100.0	-3.7e-13	0.174
GAN	86.5	52.8	7.63	0.191
VAE-GAN	100.2	54.6	7.52	0.193
Diffusion	110.8	62.1	7.60	0.193
VAE-Diffusion	113.8	64.8	6.86	0.195

The results in Table 3 show that the **VAE-Diffusion** architecture consistently outperforms other generative models across all evaluation metrics. Moreover, the diffusion-based architecture achieves better rhythmic metrics (**BCS** and **BHS**) than GAN-based approaches, while VAE-GAN exhibits slightly better **FAD** and **BAS** performance. These results confirm that diffusion models are particularly effective for audio generation tasks, and when combined with a variational autoencoder in the latent space, they produce higher-quality and more semantically consistent results.

4.6 Visualization Results

In addition to quantitative evaluation, qualitative visualization results are provided. **Figures 3 and 4** present visualization results on the **PopDanceSet** and **AIST++** datasets, respectively. In these figures, PopDanceSet results are shown in red, while AIST++ results are shown in blue. Each row represents a generated dance sequence.

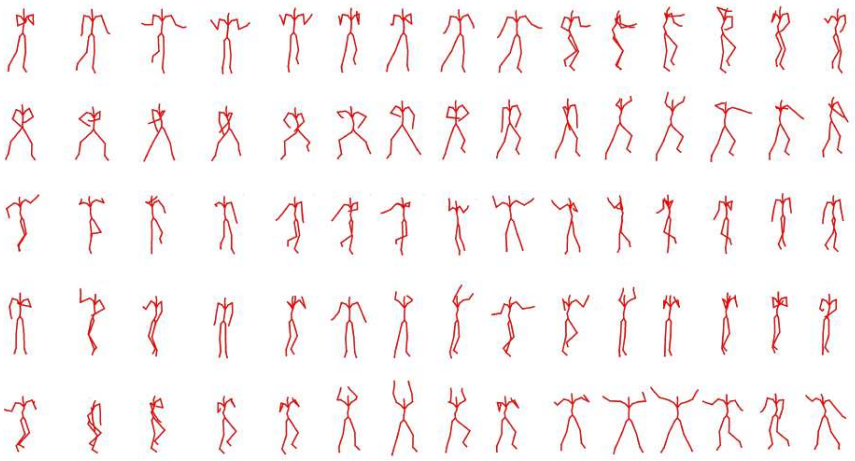


Figure 3 Visualization Results of the PopDanceSet Dataset

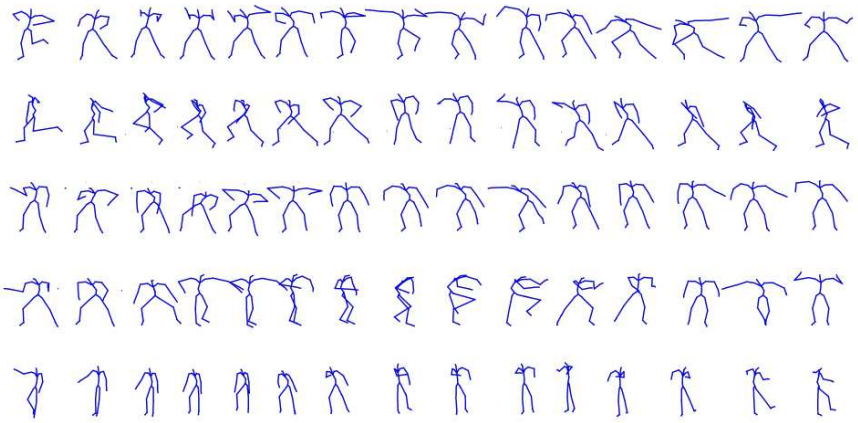


Figure 4 Visualization Results of the AIST++ Dataset

The visualization results demonstrate that the proposed method generates dance motion sequences with higher quality and greater stylistic diversity. The generated motions exhibit smooth transitions, coherent rhythmic structure, and strong alignment with musical beats, further validating the effectiveness of the proposed cycle-consistency-based framework.

5. Conclusion and Future Work

5.1 Conclusion

This paper proposed a cycle-consistency-based framework for music-driven dance motion generation and explored its application in dance teaching contexts. By constructing a closed-loop mapping between music and dance, the proposed method encourages generated dance motions to preserve meaningful musical semantics, leading to improved motion quality, motion diversity, and motion–music alignment. Experimental results on public datasets demonstrate that the proposed framework outperforms existing methods in both quantitative metrics and qualitative visualizations.

From an educational perspective, the proposed framework provides a new way to enhance the interpretability of AI-generated dance motions. By ensuring that dance movements can reconstruct the original music, the model supports a deeper understanding of how musical structure is embodied in movement. This property is particularly valuable for dance teaching, as it enables more musically grounded instructional demonstrations and helps learners better perceive rhythm, timing, and expressive cues through embodied practice.

5.2 Future Work

Future work will focus on integrating the proposed framework into real-world dance teaching systems. One promising direction is to incorporate learner-generated dance motions into the cycle-consistent framework to provide personalized feedback based on music–motion correspondence. In addition, extending the framework to support interactive teaching scenarios, such as real-time guidance and adaptive difficulty adjustment, may further enhance its educational effectiveness. Finally, user studies involving dance instructors and learners will be conducted to systematically evaluate the pedagogical impact of the proposed method in practical teaching settings.

Funding

This research received no external funding.

Author Contributions

Lun Lu conceived the idea, designed the experiments, analyzed the data, and wrote the manuscript. The author reviewed and approved the final version of the manuscript.

Data Availability

The datasets used in this study are publicly available.

The **AIST++** dataset can be accessed at: https://google.github.io/aistplusplus_dataset/

The PopDanceSet dataset can be accessed at: https://github.com/Luke-Luo1/POPDG

References

1. Kang, J., Kim, S., Lee, J., & Park, J. (2023). Dancing on the inside: A qualitative study on an online dance learning system with teacher–AI collaboration. Frontiers in Psychology, 14, 1178342.
2. Miko, H., Frizen, R., & Steinberg, C. (2025). Using AI-based feedback in dance education: A systematic literature review. Research in Dance Education.
3. Choi, J.-H., Lee, J.-J., & Nasridinov, A. (2021). Dance self-learning application and its dance pose evaluations. In Proceedings of the ACM Symposium on Applied Computing.
4. Lee, J.-J., Choi, J.-H., Chuluunsaikhan, T., & Nasridinov, A. (2020). Pose evaluation for dance learning applications using joint position and angular similarity. In Adjunct Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing.
5. Wang, Y. (2025). Enhancing remote dance education through motion capture and web-based feedback using expert reference sequences. Computers & Education: Artificial Intelligence, 6.
6. Zhou, Y., Bragg, J., Morris, M. R., et al. (2021). SyncUp: Vision-based practice support for synchronized movement skill learning. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 5(1).
7. Li, R., Yang, S., Ross, D. A., & Kanazawa, A. (2021). AI choreographer: Music conditioned 3D dance generation with AIST++. In Proceedings of the IEEE/CVF International Conference on Computer Vision.
8. Huang, R., Hu, H., Wu, W., Sawada, K., Zhang, M., & Jiang, D. (2021). Dance revolution: Long-term dance generation with music via curriculum learning. arXiv preprint.
9. Zhuang, W., Wang, C., Xia, S., & Xiong, Y. (2022). Music2Dance: DanceNet for music-driven dance generation. ACM Transactions on Multimedia Computing, Communications, and Applications, 18(3).
10. Valle-Pérez, G., et al. (2021). Transflower: Probabilistic autoregressive dance generation with multimodal attention. In Proceedings of the ACM International Conference on Multimedia.
11. Siyao, L., Yu, H., Wang, C., & Xia, S. (2022). Bailando: 3D dance generation by actor–critic GPT with choreographic memory. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
12. Gong, K., et al. (2023). TM2D: Bimodality-driven text-to-motion and music-to-dance generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
13. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., & Bermano, A. H. (2023). Human motion diffusion model. In International Conference on Learning Representations.
14. Dabral, R., et al. (2024). PriorMDM: Human motion diffusion as a generative prior. In International Conference on Learning Representations.
15. Zhang, Y., et al. (2024). OmniControl: Control any joint at any time for human motion generation. In International Conference on Learning Representations.
16. Tseng, J., Jiang, L., Liu, C., Yang, M.-H., & Yang, W. (2022). EDGE: Editable dance generation from music. arXiv preprint.

17. Luo, Z., Ren, M., et al. (2024). POPDG: Popular 3D dance generation with PopDanceSet. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
18. Alexanderson, S., Nagy, R., Beskow, J., & Henter, G. E. (2023). Listen, denoise, action! Audio-driven motion synthesis with diffusion models. ACM Transactions on Graphics, 42(4).
19. Alexanderson, S., et al. (2023). The Motorica Dance Dataset for audio-driven motion synthesis. ACM Transactions on Graphics, 42(4).
20. Zhou, Y., et al. (2024). MIDGET: Music-conditioned 3D dance generation with tokenized motion representation. arXiv preprint.
21. Fan, Z., Zhang, Z., et al. (2023). AIOZ-GDANCE: Music-driven group choreography. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
22. Wang, H., Song, Y., Jiang, W., & Wang, T. (2024). A music-driven dance generation method based on a spatial–temporal refinement model. Sensors, 24(2).
23. Zhang, C., et al. (2024). Dance2Music-Diffusion: Generating music from dance videos with latent diffusion models. EURASIP Journal on Audio, Speech, and Music Processing.
24. Chen, Y., et al. (2023). Motion-to-dance music generation using latent diffusion models. In Proceedings of the ACM International Conference on Multimedia.
25. Huang, Z., et al. (2019). Dance2Music: Automatic dance-driven music generation. In Proceedings of the ACM International Conference on Multimedia.
26. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision.