POPDG: Popular 3D Dance Generation with PopDanceSet

Zhenye Luo* Min Ren * Xuecai Hu† Yongzhen Huang†
Li Yao
School of Artificial Intelligence, Beijing Normal University
luozy2021@mail.bnu.edu.cn, {renmin, huxc1208, huangyongzhen, yaoli}@bnu.edu.cn

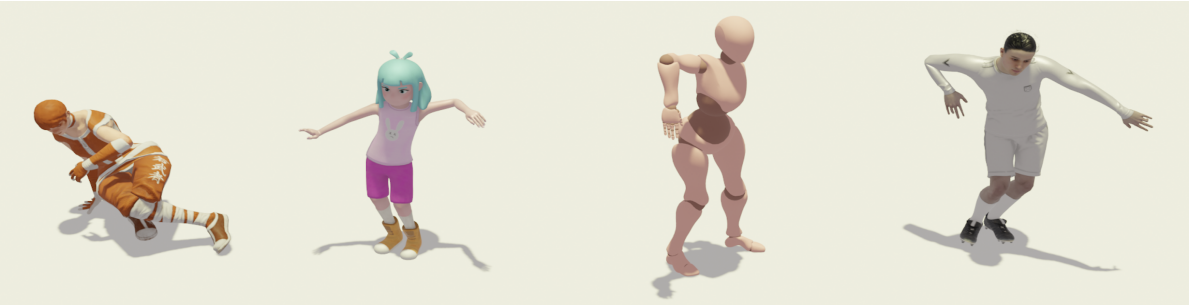

Figure 1. POPDG, in combination with PopDanceSet, could generate a variety of aesthetically driven popular dances.

Abstract

Generating dances that are both lifelike and well-aligned with music continues to be a challenging task in the cross-modal domain. This paper introduces PopDanceSet, the first dataset tailored to the preferences of young audiences, enabling the generation of aesthetically oriented dances. And it surpasses the AIST++ dataset in music genre diversity and the intricacy and depth of dance movements. Moreover, the proposed POPDG model within the iDDPM framework enhances dance diversity and, through the Space Augmentation Algorithm, strengthens spatial physical connections between human body joints, ensuring that increased diversity does not compromise generation quality. A streamlined Alignment Module is also designed to improve the temporal alignment between dance and music. Extensive experiments show that POPDG achieves SOTA results on two datasets. Furthermore, the paper also expands on current evaluation metrics. The dataset and code are available at https://github.com/Luke-Luo1/POPDG.

1. Introduction

Dance is a fundamental artistic expression with a rich history in humanity. Throughout time, humans have utilized dance to convey messages and express emotions [18].The task of music-driven dance generation not only helps choreographers improve the efficiency of creating innovative dances but also facilitates performances by virtual characters. It even extends to the field of neuroscience, assisting researchers in exploring the relationship between human movement and music [3].

This task has long been hampered by the scarcity of publicly available datasets and the limitations in generative model capabilities. As of now, the AIST++ dataset[24] is among the few with a significant volume of data that is publicly accessible. Despite significant advancements in dance generation models in recent years, issues such as the complexity of training steps, instability in generation, and lack of diversity still exist. This paper introduces the PopDanceSet and the POPDG, aimed at enhancing both the dataset and the model aspects of dance generation.

The AIST++ dataset's limitations include a lack of aesthetically oriented dances, a narrow range of dance and music genres, among others. The dances in this dataset are confined to 10 subcategories of street dance, which hardly encompass the vast array of dance styles in reality. The PopDanceSet, created through a popularity function designed in this paper, filters dance videos that align with popular aesthetics. It represents a significant breakthrough in terms of aesthetically oriented content, diversity in dance types, music genres, and dance movements.

Previous generative models primarily focused on the temporal alignment of music and dance, but even when considering the spatial constraints of dance, they did not delve

*These authors contributed equally.
†Corresponding author.

into the physical interconnections between specific joints. Instead, they approached the task holistically or attempted to learn specific movement patterns [36]. The alignment between dance and music is also crucial. Prior methods either underestimated this issue or complicated the training process [36, 45, 46]. Undoubtedly, these issues impact the overall quality and diversity of generated dances.

This paper specifically proposes a space augmentation algorithm based on Attention Mechanism, forming a dance decoder block to strengthen the spatial connections among joints in dance movements. Furthermore, a streamlined alignment module is designed to encode the spatiotemporal features of music alongside dance, thereby significantly enhancing their rhythmical synchronization.

Finally, in the task of music-driven dance generation, existing evaluation metrics have certain limitations. This paper also proposes evaluation metrics that are suited to this task, thereby enabling a more reasonable assessment of the generated dances. In summary, the contributions of this paper can be enumerated as follows:

- We build the PopDanceSet, reflecting contemporary aesthetic preferences. It significantly enriches the diversity and quantity of dances and music, increases the complexity of dance movements, and offers excellent extensibility for continuous supplementation.
- We introduce POPDG(Popular 3D Dance Generation), which is based on iDDPM and achieves a balance between generation quality and diversity. The model pays particular attention to the spatial features of the dancer's body joints, especially proposing the Space Augmentation Algorithm. In addition, our newly designed Alignment Module integrates the spatiotemporal features of music and dance, strengthening the alignment between dance and music.
- Extensive experiments were conducted in this study. It was observed that the POPDG produced the exciting results, both on AIST++ and PopDanceSet. And we also make a reasonable extension to the evaluation metrics, making the assessment of dance generation more comprehensive and objective.

2. Related Works

2.1. Music-Dance Dataset

High-quality dance generation relies on comprehensive and diverse music-dance datasets. Earlier research primarily utilized motion capture technology for limited dataset collection, as seen in [2, 39, 40, 51], or leveraged pose estimation models [4, 11, 34] to derive 2D/3D poses from online dance videos. However, due to the complexities in dance motion capture and the constraints of earlier pose estimation models, these datasets were limited in dance variety, duration, and motion capture quality. A significant advancement was made with AIST++ [24], an extension of AIST [42], offering longer durations, precise 3D joint annotations, and high-quality dance movements, setting a new standard in the field. Despite its wide usage, later databases like PMSD[43], PhantomDance[21], and MMD[5] provide only incremental advancements, mainly providing additional data for specific research tasks without much broader impact due to limited public availability.

2.2. Human dance generation

Initially, music-driven dance generation, an autoregressive task, explored the music-dance relationship using traditional machine learning algorithms [10, 20, 29], but these methods produced dances with limited duration, diversity, and poor adaptability to various melodies and rhythms. The advent of deep learning saw researchers [2, 5, 8, 15, 18, 19, 25, 31, 39, 40, 45–47, 50, 51] employing CNNs, LSTMs, MLPs, and GCNs to better capture deep features. Despite improved feature extraction and generalizability, generating dances with high diversity remains challenging. With FACT's introduction [24], Transformers have gained prominence for their superior temporal feature modeling [16, 17, 43]. Further advancements by Bailando[36, 37] and EDGE[41] using VQ-VAE, GPT, reinforcement learning, and DDPM have enhanced dance quality and diversity but at the cost of increased training complexity. The stability and overall quality of long-sequence dance generation continue to need enhancement.

2.3. Diffusion Models

Diffusion models [14], a novel class of deep generative models, learn data distributions through reverse denoising processes. They have recently shown superior generative capabilities in image generation, outperforming benchmarks in general tasks [28, 32]. Additionally, their adaptability in conditional generation tasks makes them highly versatile. Dhariwal[6] and Ho[13] demonstrated their effectiveness with guided image generation, optimizing the diversity-fidelity trade-off. Their impressive performance extends to various fields, including 3D monocular pose estimation[12] and text-driven motion generation[49]. While closely related to human pose and motion generation with emerging applications in music-driven dance generation [41], the high standards for quality and diversity in this domain mean diffusion models still necessitate further exploration.

3. PopDanceSet

3.1. Popularity Function and Dataset Construction

Our aim in building PopDanceSet was to address the issues mentioned in section 2.1 while also catering to the aesthetic preferences of contemporary youth. To this end, we devel-

Dataset	Lyrics	Aesthetics	3D $\text{Joint}_{\text{pos}}$	3D $\text{Joint}_{\text{rot}}$	2D Kpt	Genres	Subjects	Seconds
Dance with Melody[40]	✗	✗	✓	✗	✗	4	-	5640
GrooveNet[2]	✗	✗	✓	✗	✗	1	1	1380
DanceNet[51]	✗	✗	✓	✗	✗	2	2	3472
EA-MUD[39]	✗	✗	✓	✗	✗	4	-	1254
AIST++[24]	✗	✗	✓	✓	✓	10	30	**18694**
PopDanceSet(Ours)	✓	✓	✓	✓	✓	**19**	**132**	12819

Table 1. **3D Dance Datasets Comparison.** PopDanceSet stands out for its aesthetically oriented content and inclusion of music with corresponding lyrics. Encompassing a broad range of 19 genres and 132 subjects, it offers high diversity over 12,819 seconds of data, establishing itself as a valuable dataset for dance generation research.

oped a popularity function to filter suitable dance videos. We selected BiliBili[1], the video platform most popular among young people in China, as our data source. Using multiple linear regression and Student's t test [9], we identified the variables that influence video popularity, formulated the popularity function, and detailed the verification process in supplementary Sec. 7.

$$Pop = WN^T + b, \tag{1}$$

In Eq. (1), we define N as $[n_{favorites}, n_{danmucounts}, n_{views}, n_{likes}, n_{shares}]$, where each term represents the number of favorites, danmu(live comments that scroll over the video, offering an interactive and communal viewing experience) counts, views, likes, and shares respectively. These are weighted by the coefficient vector $W = [0.0251, 0.0095, 0.8033, 0.0967, 0.0243]$. Additionally, the bias term b is set to 0.0443. We establish a Pop threshold of 0.85 for selection criteria. Recognizing the inherent advantage of authors with a larger following, we consider only those videos where the view count exceeds the number of followers of the creator, denoted as $n_{views} > n_{followers}$. Moreover, we opt to exclude videos with frequent changes in camera angles or excessive shaking, to ensure data consistency and quality.

3.2. Dataset Description

We collected a total of 263 dance videos, containing 180 pieces of music. In recent years, monocular 3D joint detection technology based on SMPL[26, 48] has made significant progress, providing high-quality detection results. We employed the HybrIK model [22, 23] to extract the 3D joint features of the dancers in all videos. Each frame of data has the following parameters:

- 24 SMPL pose parameters along with the global scaling, translation and pred_scores;
- Predicted camera parameters along with root and translation;
- 17 COCO-format[33] human joint locations in 3D;

The comparison between PoPDanceSet and other datasets is in Tab. 1. It is readily apparent that PopDanceSet, while second only to the AIST++ in terms of dance duration, has comprehensively surpassed it in aspects such as dance and music genres. The collected dance genres encompass 19 categories, including CPOP, KPOP, house dance, among others, and we have endeavored to maintain an even distribution of the number of each dance type. Details of the dataset can be refer to supplementary Sec. 7. The music in this dataset spans a wide range of rhythms and styles, from classical to rock, and most retain lyrics, maintaining consistency with real-world dance environments. The complexity of movements in this dataset also exceeds that of the AIST++, meriting further research in the future.

It is particularly noteworthy that the dance data collected in the PopDanceSet consists of the pose data of human body joints for each frame and the position data in three-dimensional space, without any other parameters such as facial features or body shape. Furthermore, the accompanying music for the dances is all publicly available. Therefore, PopDanceSet does not involve any issues of privacy.

4. Method

Our model framework, as illustrated in Fig. 2, is based on iDDPM(improved-Denoising Diffusion Probabilistic Models)[28] for sampling with denoising performed by DDIM(Denoising Diffusion Implicit Models)[38]. During training, the model is fed a sequence of dance poses $x \in \mathbb{R}^{N\times 156}$spanning a certain number of frames. In line with most methods, the initial three dimensions represent the single root translation, followed by the 6-DOF (Degrees of Freedom) rotation representation of the 24 joints in the SMPL human body model. The final dimensions are binary contact labels for the feet, hands, and neck. Therefore, the pose representation is $x \in \mathbb{R}^{156=3+24\cdot 6+9}$ per frame. The model gives equal importance to music and motion. In addition to extracting 4800-dimensional music features through Jukebox[7] like EDGE, it also features a music encoder with a structure symmetrical to that of the dance decoder.

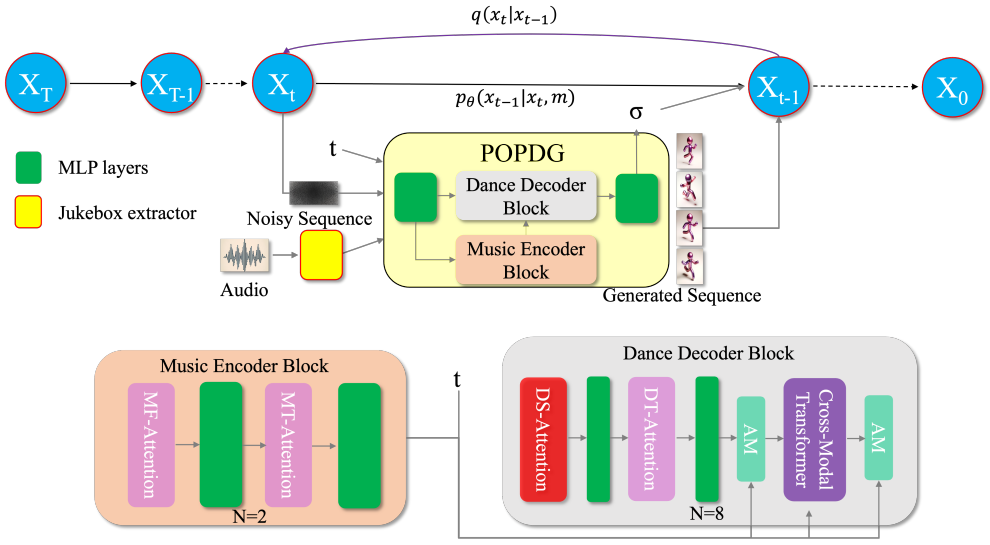

Figure 2. **POPDG Pipeline Overview.** POPDG, utilizing the iDDPM framework, learns to denoise dance sequences from time $t = T$ to $t = 0$. The audio feature sequence serves as the input to the Music Encoder Block, while the noisy sequence is input to the Dance Decoder Block, with the output being the generated dance sequence. And N refers to the stack number. Beginning with a noisy sequence $z_T \sim N(0, I)$, POPDG generates the estimated frame of the dance sequence. It then progressively noises the sequence back to $\hat{z}_T - 1$, repeating the process until $t = 0$.

4.1. Improved-DDPM and DDIM

Currently, the DDPM framework is employed in both the action domain and dance generation domain, while iDDPM, in comparison to DDPM, learns not only the mean from the data distribution but also takes variance into account, thereby increasing the diversity of generation while ensuring quality. In iDDPM, the forward process also adheres to a Markov chain $q(z_t|x)$, and we calculate the mean and variance using the following Eq. (2) and Eq. (3):

$$q(\boldsymbol{z}_t|\boldsymbol{x}) \sim \mathcal{N}(\sqrt{\bar{\alpha}_t}\boldsymbol{x}, \boldsymbol{\Sigma_t}), \tag{2}$$

$$\Sigma(\boldsymbol{x}, \boldsymbol{t}) = \exp(v \log \beta_t + (1 - v) \log \tilde{\beta}_t) \tag{3}$$

where both alpha $\bar{\alpha}_t \in (0, 1)$and β_t are hyper parameters, and the variable v is predicted by our model. In reverse process, we learn to estimate $\hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{m}) \approx \boldsymbol{x}$ with model parameter θ for all t . We optimize the basic loss as Eq. (4):

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{\boldsymbol{x},t}\left[\|\boldsymbol{x} - \hat{\boldsymbol{x}}_\theta(\boldsymbol{z}_t, t, \boldsymbol{m})\|_2^2\right] \tag{4}$$

However, this does not take into account variance. Therefore, we follow the iDDPM method, adding a variational lower bound loss. The parameters here also adhere to iDDPM, as shown in Eq. (5):

$$L_{\text{vlb}} = E_{t\sim p_t}\left[\frac{L_t}{p_t}\right], \text{where } p_t \propto \sqrt{E[L_t^2]} \text{ and } \sum p_t = 1 \tag{5}$$

Thus, the total loss of iDDPM combines Eq. (4) and Eq. (5):

$$L_{\text{hybrid}} = L_{\text{simple}} + \lambda L_{\text{vlb}} \tag{6}$$

For the denoising process, we follow the DDIM method. This method allows for significantly faster training and inference without much compromise on the quality of the generation.

4.2. Music and Dance Spatiotemporal block

As shown in Fig. 2, POPDG comprises two blocks: the music encoder and the dance decoder. These two blocks, based on the principle of symmetry and empirical validation, have similar spatiotemporal Transformer modules. The core of two blocks lies in four attention mechanisms: MF-Attention (Music Feature-Attention), MT-Attention (Music Temporal-Attention), DS-Attention (Dance Spatial-Attention) and DT-Attention (Dance Temporal-Attention).

4.2.1 Dance Decoder Block

The details of the dance decoder block can be found in Fig. 2, which is composed of Transformer based on DS-Attention and DT-Attention. Previous methods primarily used DT-Attention. The input dance sequence is x_{motion}. We take the positional encoded x_{motion} as Q, K and the original x_{motion} as V, and pass through the classic attention[44]:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}, \mathbf{M}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T + \mathbf{M}}{\sqrt{C}}\right)\mathbf{V} \tag{7}$$

where $Q * K^T$ results in attention map with $[N \times N]$. M represents Mask operation. We randomly mask some frames in

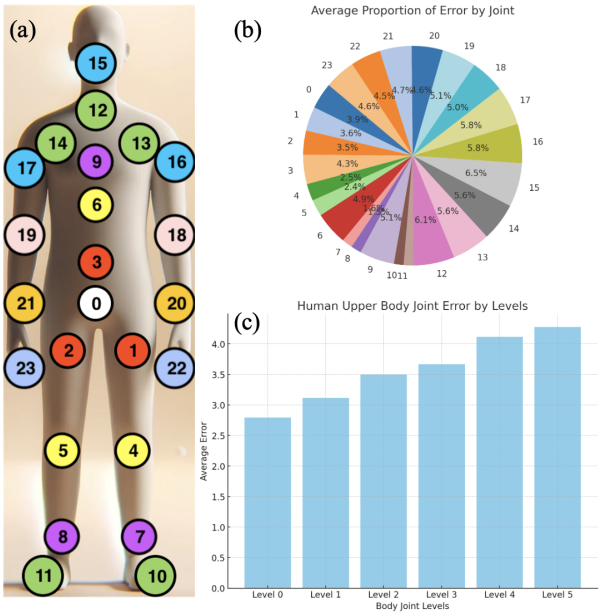

Figure 3. **Analysis of Joint Error Distribution in SMPL Human Body Model.** (a) SMPL Joint Labeling: Marks human body joints from the hip (level 0 joint) outward, color-coded by different levels. (b) Joint Error Proportions: Shows that upper body joints experience increasing error the further they are from the hip. (c) Upper Body Joint Error Levels: Displays average errors across upper body joint levels.

the dance sequence to enhance robustness. It mainly pays attention to the temporal relationship within the input sequence.

In POPDG, we place a spatial attention, DS-Attention, to capture the spatial connections between human body joints, as illustrated in Fig. 4. In the model, $x_{motion} \in \mathbb{R}^{b \times N \times h}$, where h represents the hidden feature dimension of the dance posture. We transpose x_{motion} before feeding it into DS-Attention, thereby obtaining an attention map focused on the spatial dimension. The additional Space Augmentation Algorithm within DS-Attention is capable of capturing the actual spatial connections between the joints.

SMPL designates the hip as the root joint, with other body joints categorized into levels based on their distance from the hip. As shown in Fig. 3(a), joints with the same background color are at the same level. Comparing generated dance movements with ground truth data, we observe in Fig. 3(b) that joint errors increase with distance from the root joint. The average error spans from 4% at the root to approximately 6.5% at the upper body parts like the ribs and neck. Fig. 3(c) displays the average error across different joint levels.

In the traditional multi-head attention model, attention weights are calculated based on the similarity between queries and keys. To better capture the spatial relationship

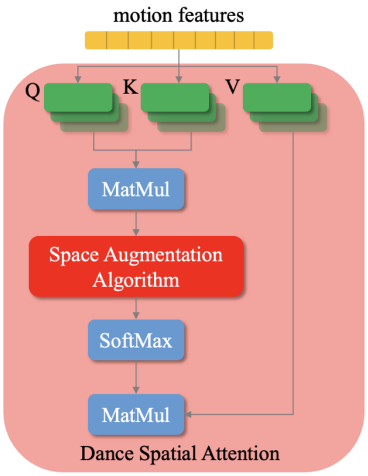

Figure 4. **The Overview of Dance Spatial Attention.** The key distinction between dance spatial attention and standard multi-head attention is the incorporation of the Space Augmentation Algorithm when calculating the Attention Map between Query and Key. This algorithm is tailored to emphasize the upper body joints in relation to the hip, enhancing their spatial inter connectivity.

between specific joints, we introduced a **Space Augmentation Algorithm**. This algorithm enhances weights based on the distance of the joints from the root joint.

The physical meaning of the algorithm is to strengthen the relationship between each joint in the upper body and its parent joint. Specifically, assuming the $(i+1)^{th}$ joint m is above the i^{th} joint n. It is known that in the calculation of the $attentionmap \in \mathbb{R}^{[24,24]}$ in DS-Attention, we already have $(0, m), (m, 0)$ and $(m, n), (n, m)$ which represent the connection weights of joint m with the root joint and joint n, respectively. If we add the value of $(0, n)$ to $(0, m)$ and $(m, 0)$, essentially we enhance the connection between joint m and the root joint. Similarly, we can continuously pass information from parent joints to downstream-level joints. The specific algorithm implementation can be summarized by Algorithm 1. From an experimental perspective, it is viable with or without dividing by 2.

Algorithm 1 Space Augmentation Algorithm

function APPLYWEIGHTING($attn\_probs$)
  $levels \leftarrow \{0 : [3], 3 : [6], 6 : [9], 9 : [12, 13, 14], 12 : [15], 13 : [16], 14 : [17]\}$
  for $source, targets$ **in** $levels$ **do**
    for $target$ **in** $targets$ **do**
      ENHANCE($attn\_probs, source, target$)
  return $attn\_probs$
function ENHANCE($attn, src, tgt$)
  $attn[0, tgt] \mathrel{+}= attn[0, src]; \ (attn[0, tgt] \ /= 2)$
  $attn[tgt, 0] \mathrel{+}= attn[src, 0]; \ (attn[tgt, 0] \ /= 2)$

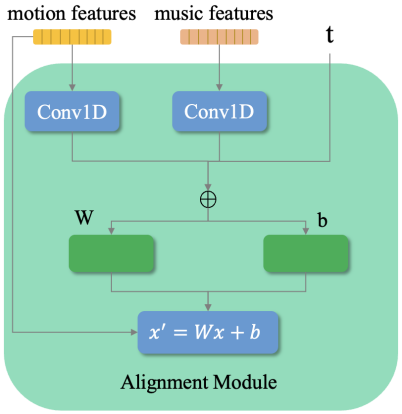

Figure 5. **The overview of Alignment Module:** Once the music and dance features have been processed through temporal and spatial Transformers, we apply temporal feature processing to both.

4.2.2 Music Encoder Block

Following the principle of symmetry, we adopted a design for the music encoder block that mirrors that of the dance decoder block, and the ablation experiments are displayed in Section 5.4. Building on the existing MT-Attention, we transpose the music data. Unlike dance motions with clear temporal and spatial definitions, after passing through MF-Attention, the musical feature x_{music} will obtain relationships between mathematical features such as MFCC and chroma.

4.3. Alignment Module

The quality of generated dance is also contingent on its compatibility with the music. Therefore, building upon the work in Section 4.2, we designed a concise alignment module that can enhance the adaptability of dance to music while ensuring the quality of dance generation.

Before feeding the dance and music data into our module, we apply temporal processing to both. Unlike previous methods generally applied spatial position encoding to dance sequences, our work equally emphasizes both temporal and spatial characteristics of dance. This involves performing a one-dimensional convolution operation on both sets of features and adding the resultant values to the time step t in the diffusion model. These combined features are then fed through MLP, consistent with DenseFiLM[30]. The detailed model structure is depicted in Fig. 5.

4.4. Loss Function

In the training process, in addition to the initial loss function given by Eq. (6), we integrated and promoted the training strategies of previous methods. Additional information is in supplementary Sec. 8:

- velocity and acceleration loss:

$$\mathcal{L}_{\text{va}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(\boldsymbol{x}^{'} - \hat{\boldsymbol{x}}^{'})\|_2^2 + \|(\boldsymbol{x}^{''} - \hat{\boldsymbol{x}}^{''})\|_2^2 \quad (8)$$

  We calculate the average error of speed and acceleration between generated dance x and real dance $\hat{x}$.

- FK loss and Body Loss:

$$\mathcal{L}_{\text{body}} = \frac{1}{N-1}\sum_{i=1}^{N-1} \|(FK(\hat{\boldsymbol{x}}^{(i+1)}) - FK(\hat{\boldsymbol{x}}^{(i)})) \cdot \hat{\boldsymbol{b}}^{(i)}\|_2^2 \quad (9)$$

  We adopt the same FK loss function as used in the EDGE. While L_{body} upgrades $L_{contact}$, extending its focus from solely the feet to include the hands and the neck. FK$(\cdot)$ denotes the forward kinematic function that converts joint angles into joint positions. $\hat{b}^{(i)}$ is the model's own prediction of the binary body contact label's portion of the pose at each frame i.

Finally, by combining these loss functions, we formulate the loss function used for training POPDG, as Eq. (10).

$$\mathcal{L} = \mathcal{L}_{\text{hybrid}} + \lambda_{\text{FK}}\mathcal{L}_{\text{FK}} + \lambda_{\text{va}}\mathcal{L}_{\text{va}} + \lambda_{\text{body}}\mathcal{L}_{\text{body}} \quad (10)$$

5. Experiments

5.1. Implement Details

In this study, the generative capabilities of the POPDG model are demonstrated on PopDanceSet and AIST++. Initially, to ensure the intrinsic generation quality of the dataset, the construction process included manual checks to confirm the reliability of the extracted dance generation quality. For the PopDanceSet training, there were 736 video segments utilized as the training set and 24 video segments used for testing, ensuring that the dances and music in the test set had not appeared in the training set. The experimental procedure on the AIST++ mirrored the previous methodology, with the training and test sets comprising 952 and 40 videos, respectively, and generating dance sequences lasting 25 seconds in duration.

The entire experimental process took around 100 hours on two A800 GPUs for the AIST++ and 66 hours for the PopDanceSet. The dance decoder block's parameter settings were similar to those used for 3D pose estimation, with the hidden layer dimension uniformly at 512, and MLP layers dimension at 1024. These two parameters were consistently applied in the music encoder block. DT-Attention, MF-Attention, and MT-Attention all employed the conventional 8-head attention mechanism. The optimizer chosen for the model was Adan, with a learning rate set at 0.001 and betas of 0.02, 0.08, and 0.01, with an eps of 1e-8.

5.2. Evaluation Metrics

5.2.1 Motion Quality

Researchers commonly employ the FID (Frechet Inception Distance) [27] metric to assess the motion quality of generated dances. However, experiments often reveal that, despite some dances scoring well on FID, they exhibit poor visual quality. In response, EDGE introduced the Physical Foot Contact (PFC) score, denoted by Eq. (11), which assesses the plausibility of dance movements directly through the acceleration of the hips and the velocity of the feet. But since PFC only considers the lower body and dance is a full-body movement, it is also important to consider the upper body[35]. Therefore, this paper builds upon PFC by including the neck and hands, extending the evaluation to the full body to create the PBC (Physical Body Contact) score.

$$f(x,y,z) = \frac{\sum_{i=1}^{N} ||\overline{\boldsymbol{a}}_x^i|| \cdot ||\mathbf{v}_y^i|| \cdot ||\mathbf{v}_z^i||}{\max_{1\le j\le N} ||\overline{a}_x^j||} \tag{11}$$

$$PBC = \frac{1}{N}\left[-f(\text{root}, \text{lfoot}, \text{rfoot}) + f(\text{lchest}, \text{lhand}, \text{null}) + f(\text{rchest}, \text{rhand}, \text{null}) + f(\text{neck}, \text{head}, \text{null})\right] \tag{12}$$

In Eq. (12), the variables $||\overline{\boldsymbol{a}}^j_{\text{root}}||$, $||\overline{\boldsymbol{a}}^j_{\text{lchest}}||$, $||\overline{\boldsymbol{a}}^j_{\text{rchest}}||$ and $||\overline{\boldsymbol{a}}^j_{\text{neck}}||$ each represent the average acceleration of the root joint, the left and right chest joints, and the neck joint of the SMPL model, respectively, projected onto the XYZ plane for each frame i. Compared to PFC, PBC incorporates a broader consideration of the plausibility of dance movements. For detailed elaboration, please refer to supplementary Sec. 9.

5.2.2 Motion Diversity

In terms of the diversity of generated dances, we have adopted the Div_k and Div_g metrics used by previous methods, which measure the average kinematic and geometric distance between generated dances and the ground truth to quantify diversity.

5.2.3 Motion-Music Correlation

The match between music and dance is also a crucial factor affecting the quality of generated dances. We also employ the formula from previous models to measure the synchrony between dance and music. The Beat Alignment Score adopted follows FACT, defined as:

$$BeatAlign = \frac{1}{|B^m|}\sum_{t^m\in B^m} \exp\left\{-\frac{\min_{t^d\in B^d}\|t^d - t^m\|^2}{2\sigma^2}\right\} \tag{13}$$

where B^m and B^d record the time of beats in dance and music, respectively. And σ is normalized parameter which is set to be 3 in our experiment.

5.3. Comparing to Existing Methods

As illustrated in Tab. 2, a comparison with existing methods reveals that our experiment demonstrates the superiority of the POPDG over previous models on the AIST++ and PopDanceSet. Specifically, in the PopDanceSet experiment, POPDG outperformed all other methods, achieving the most optimal results. In terms of the PFC and PBC metrics, POPDG surpassed EDGE by 1.6004(26.8%) and 0.4672(7.98%) in motion quality, respectively. Moreover, POPDG also excelled in dance generation diversity, as evidenced by its superior performance on the Div_k and Div_g metrics, where it improved by 1.2576(34.9%) and 0.2878(5.02%) compared to EDGE. The enhancement in diversity can be attributed to the iDDPM, as discussed in Section 5.4, which augments dance diversity by predicting the mean and variance of motion data. Furthermore, in the BAS metric, POPDG also surpassed Bailando, which specifically employed reinforcement learning to enhance this aspect. Also demonstrated in Section 5.4, the AM strengthens the alignment between music and dance. On the AIST++, although POPDG did not exhibit as significant an impact as in the PopDanceSet, it still surpassed previous works in most evaluation metrics.

5.4. Ablation Studies

- **Modules in POPDG** Tab. 3 shows the impact of incorporating MF-Attention, DS-Attention, and AM on the quality and music alignment of generated dances. By strengthening the connections between the dancer's body joints, we've improved the quality of the generated dances. Adding MF-Attention has also enhanced the outcomes, considering the model's symmetry. While the exact relationships between various mathematical features of music are still unclear, we believe that POPDG has captured their deeper correlations. The newly designed AM has achieved positive results in both dance quality and alignment, likely due to our modulation of dance using the temporal features of both music and dance.
- **iDDPM** The fundamental difference between the DDPM and iDDPM generative frameworks lies in the fact that while DDPM predicts the mean of the generated data, iDDPM takes into account the variance as well. Theoretically, this gives iDDPM a stronger generative capacity compared to DDPM. For the task of music-driven dance generation, there has traditionally been a trade-off between generation quality and diversity, where improvements in the quality of generated dances tend to reduce diversity. However, the use of iDDPM has allowed us to achieve a favorable balance between generation quality and diversity. This is substantiated by the results presented in Tab. 4.

Dataset	Method	Motion Quality		Motion Diversity		Motion-Music Corr
		PFC ↓	PBC →	Div_k ↑	Div_g ↑	Beat Align Scores ↑
PopDanceSet	GroundTruth	1.2302	2.8485	6.4034	7.0289	0.330
	FACT	7.5663	8.1007	3.7371	<u>5.7843</u>	0.405
	Bailando	6.1762	5.9237	4.2253	5.5396	0.480
	EDGE	5.9701	5.8535	3.6065	5.7350	0.475
	POPDG	**4.3697**	**5.3863**	**4.8641**	**6.0228**	**0.482**
AIST++	Ground Truth	0.3152	3.6231	8.6916	7.5159	0.510
	FACT	1.1722	8.6751	7.5213	<u>6.6993</u>	0.422
	Bailando	0.9268	6.8409	6.2411	**5.7120**	0.467
	EDGE	0.9201	6.5191	6.1040	3.1415	0.456
	POPDG	**0.8014**	**6.2419**	**7.5374**	3.6707	**0.469**

Table 2. **Dance Quality Comparison on the PopDanceSet and AIST++ Test Sets.** For PopDanceSet, we repused the Bailando and EDGE models, and specifically developed a PyTorch version of FACT to generate dances matching POPDG in length. On AIST++, we continued using the top three previous models for comparison. POPDG mostly outperformed others in motion quality, diversity, and music-dance alignment on both datasets. Notably, the high Div_g scores of FACT are skewed by excessive, meaningless swaying in both datasets. In our metrics, ↑ signifies that higher values indicate better performance, ↓ implies the opposite, and → represents that values closer to the Ground Truth are better.

baseline	MF-Attn	DS-Attn	AM	PFC ↓	PBC →	BAS ↑
✓				0.9600	5.5712	-
✓	✓			0.9290	5.3363	-
✓	✓	✓		0.9116	5.0391	0.431
✓	✓	✓	✓	**0.8487**	**4.9626**	**0.448**

Table 3. **Modules in POPDG Ablation Study Results.** The ablation study, conducted under the constraints of experimental conditions on an NVIDIA 3090 with the model at half dimension, demonstrates the positive impact of each proposed module on dance generation. Significant improvements were observed in PFC, PBC, and BAS metrics as additional modules were integrated.

Method	PFC ↓	PBC →	Div_k ↑	Div_g ↑
POPDG	**0.8014**	6.2419	7.5374	**3.6707**
w/o iDDPM	1.1077	**6.1515**	**8.3189**	3.3699

Table 4. **iDDPM Ablation Study Results.**This experiment validates that the iDDPM framework effectively balances motion quality and diversity.

5.5. User Study

Our user study involved twenty participants. All participants in the user study were between the ages of 23 and 25. We first trained POPDG on PopDanceSet and AIST++, then selected ten segments of wild music as input, and provided the model-generated dance pairings for participants to evaluate. Participants chose which dance segment they found more appealing. As Tab. 5 indicates, the majority(70%) preferred the dances from PopDanceSet, demonstrating that PopDanceSet successfully caters to the aesthetic preferences of the young people. The specific visual rendering effects can be referred to in the supplementary Sec. 10.

Dataset	User Study
	Our Dataset Wins
PopDanceSet	-
AIST++	70.0% ± 20.0%

Table 5. **User Study Results.**The comparative study between PopDanceSet and AIST++ demonstrates a clear preference for our database among modern youth.

6. Conclusion and Discussion

In this study, we introduce PopDanceSet to enrich data in the field, increase the complexity of dance movements, and reflect contemporary aesthetics. The POPDG model introduced in this paper enhances the connectivity among the dancer's body parts through DS-Attention and improves the alignment between the generated dance and music using AM. Its iDDPM framework maintains a balance between dance quality and diversity, achieving great results on both PopDanceSet and AIST++. While the model is trained end-to-end, the training cost remains relatively high. Future research should explore strategies to balance the diversity and quality of generated dances using more lightweight models. Additionally, developing metrics that can objectively evaluate dance quality is also crucial for the task. And we also believe that a deeper study of music deserves our continued effort.

References

[1] BiliBili. https://www.bilibili.com/. 3, 1

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

[3] Steven Brown and Lawrence M Parsons. The neuroscience of dance. *Scientific American*, 299(1):78–83, 2008. 1

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[5] Kang Chen, Zhipeng Tan, Jin Lei, Song-Hai Zhang, Yuan-Chen Guo, Weidong Zhang, and Shi-Min Hu. Choreomaster: choreography-oriented music-driven dance synthesis. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2

[7] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. 3

[8] Yinglin Duan, Tianyang Shi, Zhipeng Hu, Zhengxia Zou, Changjie Fan, Yi Yuan, and Xi Li. Automatic translation of music-to-dance for in-game characters. In *IJCAI*, pages 2344–2351, 2021. 2

[9] Lynn E Eberly. Multiple linear regression. *Topics in Biostatistics*, pages 165–187, 2007. 3

[10] Rukun Fan, Songhua Xu, and Weidong Geng. Example-based automatic music-driven conventional dance motion synthesis. *IEEE transactions on visualization and computer graphics*, 18(3):501–515, 2011. 2

[11] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[12] Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, and Jun Liu. Diffpose: Toward more reliable 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13041–13051, 2023. 2

[13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Ruozi Huang, Huang Hu, Wei Wu, Kei Sawada, Mi Zhang, and Daxin Jiang. Dance revolution: Long-term dance generation with music via curriculum learning. *arXiv preprint arXiv:2006.06119*, 2020. 2

[16] Yuhang Huang, Junjie Zhang, Shuyan Liu, Qian Bao, Dan Zeng, Zhineng Chen, and Wu Liu. Genre-conditioned long-term 3d dance generation driven by music. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4858–4862. IEEE, 2022. 2

[17] Jinwoo Kim, Heeseok Oh, Seongjean Kim, Hoseok Tong, and Sanghoon Lee. A brand new dance partner: Music-conditioned pluralistic dancing controlled by multiple dance genres. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3490–3500, 2022. 2

[18] Kimerer LaMothe. The dancing species: how moving together in time helps make us human. *Aeon, June*, 1:1, 2019. 1, 2

[19] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. Dancing to music. *Advances in neural information processing systems*, 32, 2019. 2

[20] Minho Lee, Kyogu Lee, and Jaeheung Park. Music similarity-based approach to generating dance motion sequence. *Multimedia tools and applications*, 62:895–912, 2013. 2

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

[22] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3383–3393, 2021. 3

[23] Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023. 3

[24] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13401–13412, 2021. 1, 2, 3

[25] Weipeng Li, Boyuan Ren, Haoyue Xu, Shiyuan Cao, and Yuyangsong Xie. Autodance: Music driven dance generation. In *2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM)*, pages 55–59. IEEE, 2021. 2

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023. 3

[27] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 7

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2, 3

[29] Ferda Ofli, Engin Erzin, Yücel Yemez, and A Murat Tekalp. Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis. *IEEE Transactions on Multimedia*, 14(3):747–759, 2011. 2

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[31] Xuanchi Ren, Haoran Li, Zijian Huang, and Qifeng Chen. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 46–54, 2020. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[33] Matteo Ruggero Ronchi and Pietro Perona. Benchmarking and error diagnosis in multi-instance pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 369–378, 2017. 3

[34] Arindam Sengupta, Feng Jin, Renyuan Zhang, and Siyang Cao. mm-pose: Real-time human skeletal posture estimation using mmwave radars and cnns. *IEEE Sensors Journal*, 20 (17):10032–10044, 2020. 2

[35] Nao Shikanai, Worawat Choensawat, and Kozaburo Hachimura. Movement characteristics of entire bodies in dancers' interaction. In *2014 14th International Conference on Control, Automation and Systems (ICCAS 2014)*, pages 1357–1361. IEEE, 2014. 7

[36] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando: 3d dance generation by actor-critic gpt with choreographic memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11050–11059, 2022. 2

[37] Li Siyao, Weijiang Yu, Tianpei Gu, Chunze Lin, Quan Wang, Chen Qian, Chen Change Loy, and Ziwei Liu. Bailando++: 3d dance gpt with choreographic memory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[38] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

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

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

[41] Jonathan Tseng, Rodrigo Castellon, and Karen Liu. Edge: Editable dance generation from music. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 448–458, 2023. 2

[42] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *ISMIR*, page 6, 2019. 2

[43] Guillermo Valle-Pérez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower: probabilistic autoregressive dance generation with multimodal attention. *ACM Transactions on Graphics (TOG)*, 40(6):1–14, 2021. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[45] Shuang Wu, Zhenguang Liu, Shijian Lu, and Li Cheng. Dual learning music composition and dance choreography. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3746–3754, 2021. 2

[46] Shuang Wu, Shijian Lu, and Li Cheng. Music-to-dance generation with optimal transport. *arXiv preprint arXiv:2112.01806*, 2021. 2

[47] Zijie Ye, Haozhe Wu, Jia Jia, Yaohua Bu, Wei Chen, Fanbo Meng, and Yanfeng Wang. Choreonet: Towards music to dance synthesis with choreographic action unit. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 744–752, 2020. 2

[48] Jinlu Zhang, Zhigang Tu, Jianyu Yang, Yujin Chen, and Junsong Yuan. Mixste: Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13232–13242, 2022. 3

[49] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*, 2022. 2

[50] Aihua Zheng, Menglan Hu, Bo Jiang, Yan Huang, Yan Yan, and Bin Luo. Adversarial-metric learning for audio-visual cross-modal matching. *IEEE Transactions on Multimedia*, 24:338–351, 2021. 2

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