DGFM: Full Body Dance Generation Driven by Music Foundation Models

Xinran Liu[1]**, Zhenhua Feng**[2]**, Diptesh Kanojia**[1]**, Wenwu Wang**[1]
[1]University of Surrey, UK, [2]Jiangnan University, China

Abstract

In music-driven dance motion generation, most existing methods use hand-crafted features and neglect that music foundation models have profoundly impacted cross-modal content generation. To bridge this gap, we propose a diffusion-based method that generates dance movements conditioned on text and music. Our approach extracts music features by combining high-level features obtained by music foundation model with hand-crafted features, thereby enhancing the quality of generated dance sequences. This method effectively leverages the advantages of high-level semantic information and low-level temporal details to improve the model's capability in music feature understanding. To show the merits of the proposed method, we compare it with four music foundation models and two sets of hand-crafted music features. The results demonstrate that our method obtains the most realistic dance sequences and achieves the best match with the input music.

1 Introduction

Music-driven dance video generation is a challenging task in cross-modal content generation [1]. The complexity comes from the need to generate dance movements that follow choreographic rules and are aligned precisely with the music. Specifically, the generated motions must be expressive while corresponding to rhythm, melody, and other musical elements. This requires the generative model to have outstanding performance in music feature extraction.

Early music-driven dance generation approaches frame the task as a similarity-based retrieval problem [2–5] due to limited training data. These methods significantly restrict the diversity and creativity of the outputs. With the advent of deep learning, recent advancements have treated music-driven dance generation as a generative task [6–10]. However, most existing models are trained on datasets with limited joint representations (*e.g.*, 24 joints), neglecting finer hand motion details, which are essential for enhancing the realism and expressiveness of the generated dances. Furthermore, the existing approaches rely on hand-crafted musical features such as Mel-Frequency Cepstral Coefficient (MFCC) [11], chroma, or one-hot beat features, which cannot fully capture the intricate link between music and dance movements [12]. In contrast, recent advances in music foundation models have demonstrated the potential of modern machine learning techniques to better understand and process music in more sophisticated ways [13]. These models trained on large-scale music datasets with minimal supervision, serve as the foundation for multiple derived models capable of performing a wide range of tasks, including music classification [14], music understanding [15, 16], and cross-model generation [17, 18]. The potential of music foundation models in music-driven dance generation remains under-explored, making it important to investigate how these models can contribute to the quality and expressiveness of the dance generation task.

In this paper, we propose a diffusion-based method, namely Dance Generation driven by the Music Foundation Models (DGFM), which enhances the quality of the generated dance movements by incorporating the classical hand-crafted audio features with high-level features obtained by the pre-trained large music foundation model. Specifically, we use Wav2CLIP [19] for high-level music

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

feature extraction. Wav2CLIP is an audio-visual model trained by distilling the knowledge from Contrastive Language-Image Pretraining (CLIP) [20]. Unlike the models that rely solely on text and audio inputs, Wav2CLIP benefits from this additional visual context, enabling more comprehensive learning and significantly improving motion prediction. Additionally, we use Short-Time Fourier Transform (STFT) as hand-crafted audio features and employ CLIP to extract features from genre prompts. By building on these components, our method captures a deeper understanding of the relationship between music and movement, resulting in realistic and intricately detailed 3D dance motions.

The contributions of our work can be summarized as follows: (1) We introduce DGFM which integrates both music and text features as inputs. It improves hand-crafted audio features by incorporating Wav2CLIP, significantly enhancing the dance generation quality. (2) To investigate the impact of music understanding on dance motion generation, we compare different music foundation models and hand-crafted music features. The results demonstrate that the combination of Wav2CLIP with STFT features achieves promising results. (3) Extensive experiments performed on the FineDance dataset demonstrate the effectiveness of our approach.

2 Related work

Music Foundation Model: Music foundation models are pre-trained on large-scale music datasets, which are designed to gain a deeper understanding of underlying musical structures, genres, and instruments [21]. The existing music foundation models can be divided into two categories. In the first category, the models are pre-trained with a single modality. This category includes Wav2Vec 2.0 [22], which is a self-supervised model that learns audio representations from raw audio through contrastive learning. Additionally, Li *et al.* [13] proposed MERT, a model for music understanding that takes advantage of Residual VQ-VAE (RVQ-VAE) and teacher models to extract musical features, facilitating pre-training based on mask language modeling (MLM). Jukebox [23] compresses raw audio into discrete codes using a multi-scale VQ-VAE and models them with an autoregressive Transformer, applied in [24] to enhance previous hand-crafted audio feature extraction strategies for dance generation task. In the second category, the models are pre-trained by multi-modal data, such as CLAP [25] which leverages the latent space derived from both audio and text to develop continuous audio representations. Building on this, AudioLDM [26] utilizes CLAP to train Latent Diffusion Models (LDMs) with audio embeddings, while text embeddings are used as conditions during sampling. Wu *et al.* [19] proposed Wav2CLIP based on CLIP [20], which projects audio into a shared embedding space along with images and text to pretrain the audio encoder. However, the application of these models in music-driven dance generation remains under-explored. It is important to investigate how music foundation models can contribute to enhancing dance generation tasks.

Music Driven Dance Generation: There has been extensive research exploring music-conditioned dance generation. Early studies [2–4] formulate this task as a similarity-based retrieval problem due to limited training data, which substantially limits the diversity and creativity of the generated dance motions. With the development of deep learning, the mainstream approaches synthesize dance segments from scratch via motion prediction, with methods such as convolutional neural network (CNN) [27, 28], recurrent neural network (RNN) [29–31], and Transformer [8, 10, 32]. However, these frame-by-frame prediction models frequently face challenges like error accumulation and motion freezing [33]. Recent studies have focused on framing the task as a generative pipeline. Built on VQ-VAE, TM2D [34] integrates both music and text instructions to generate dance movements that are consistent with the provided music and contain semantic information. Bailando [6] summarizes meaningful dance units into a quantized codebook and incorporates a reinforcement-learning-based evaluator to ensure alignment between beats and movement during dance generation. EDGE [24] apply a diffusion-based dance generation model, and also introduce a novel evaluation method based on human physical plausibility. However, all these models are trained on datasets with 24 body joints and neglect the quality of the hand motions generated. To mitigate this issue, Li *et al.* [35] proposed FineNet and introduced a new dataset with 52 joints. Despite these developments, almost all the existing models rely on hand-crafted musical features, which may lack an understanding of the connection between music and dance movements.

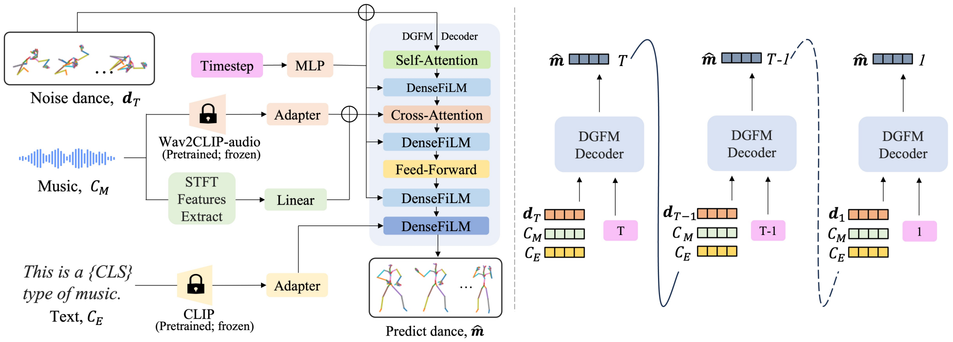

Figure 1: An overview of the proposed DGFM method.

3 Proposed Approach

In this section, we introduce our approach for generating dance movements conditioned on both music sequences and text. Given a long music piece, we first split it into N segments with length k and extract the 2D music feature map $M \in \mathbb{R}^{k \times D}$, where D represents the dimension of music features. Our objective is to generate N dance clips of length k or a long dance movement.

The existing approaches usually neglect the importance of music feature representation. Therefore, we present to incorporate Wav2CLIP [19] to our music encoder, which is an audio-visual correspondence model that distills from the CLIP framework. It is trained on VGGSound [36], a YouTube audio-visual video dataset containing approximately 200k audio clips of length 10 seconds (16kHz sampling rate) labeled with 309 classes. For hand-crafted music features, we utilize the Librosa toolbox [37] to extract STFT features. Additionally, for music genre labeling, we apply a prompt learning method [38] to expand the label into a full sentence. For *e.g.*, given the genre label "Jazz," the sentence generated is "This is a Jazz type of music." We then use CLIP [20] to extract features from this sentence.

3.1 Preliminaries of Diffusion Models

We use a diffusion model for dance generation, which has two stages: the diffusion process and the reverse process. In the diffusion process, we follow the approach outlined in Denoising Diffusion Probabilistic Models (DDPM) [39]. It defines a Markov chain that progressively adds Gaussian noise to the ground truth data $\boldsymbol{m}_0$, while allowing for the sampling of $\boldsymbol{d}_t$ at any arbitrary timestep t:

$$q\left(\boldsymbol{d}_t \mid \boldsymbol{m}_0\right)=\mathcal{N}\left(\boldsymbol{d}_t ; \sqrt{\bar{\alpha}_t} \boldsymbol{m}_0,\left(1-\bar{\alpha}_t\right) \boldsymbol{I}\right) \tag{1}$$

where $\bar{\alpha}_t$ is a constant within the range $(0,1)$ that is monotonically decreasing. As t increases and $\bar{a}_t$ approaches 0, the distribution of $\boldsymbol{d}_t$ converges to the standard normal distribution $\mathcal{N}(0, \boldsymbol{I})$. We use $T=1000$ timesteps in our model.

In the denoising process, we follow EDGE [24] and develop an attention-based network f_{rev} to reverse the forward diffusion process. The music features C_M and text features C_E as used as the input conditions to predict the movement of the dance for all t. We adopt the loss function $\mathcal{L}_S$ in DDPM as our objective, optimizing it by learning to estimate $f_{rev}(\boldsymbol{d}_t, t, C_M, C_E) \approx \boldsymbol{m}_0$, where the model refines the noisy latent variable to approximate the true data distribution. Therefore, the training objective can be defined as:

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

3.2 The Proposed DGFM Method

In this section, we introduce DGFM, a diffusion-based model that utilizes both text and audio inputs to generate full body dance movements while clearly reflecting a distinct dance genre. The overall architecture of DGDM is illustrated in Figure 1. The input music is first divided into N 4-second segments $\left\{\hat{C}_M^i\right\}_{i=1}^N$. Next, we extract the 512-dimensional Wav2CLIP features $\hat{C}_{FM}^i \in \mathbb{R}^{T \times 512}$ and the 193-dimensional STFT features $\hat{C}_{STFT}^i \in \mathbb{R}^{T \times 193}$, and feed them into an adapter and a linear layer, respectively. The linear layer projects the STFT features to a 512-dimensional vector, then we combine them through the addition operation to obtain $C_M \in \mathbb{R}^{T\times 512}$. For the input music genre label, we apply CLIP to extract 512-dimensional features $C_E \in \mathbb{R}^{512}$.

Following EDGE, the input pose data are represented as $\boldsymbol{m} \in \mathbb{R}^{k\times 319}$, according to the Skinned Multi-Person Linear (SMPL) format [40]. This representation consists of three components: a 4-dimensional foot-ground contact binary label, a 3-dimensional root translation, and 312-dimensional rotation information using a 6-dimensional rotation representation. At each denoising timestep t, DGFM predicts the denoised result and reintroduces noise from timestep $t-1$ down to 0. We repeat this process until all poses are generated at timestep 0. We feed the motion and music features into a Transformer-based denoising network, which consists of a self-attention module, a cross-attention module, multilayer perceptrons, and time-embedding Feature-wise Linear Modulation (FiLM) [41]. Subsequently, a text-specific FiLM module is incorporated, taking the output from the previous layer Y and the text embedding C_E as inputs. The embeddings are processed as follows:

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

where α is a text embedding adapter used to adjust the embedding representation. θ_w and θ_b represent the linear projections responsible for computing the weights and biases, respectively. Apart from the reconstruction loss in Equ. 2, we incorporate several auxiliary losses frequently used in motion generation tasks to improve the training stability and physical fidelity. Similarly to previous studies [42], we use the forward kinematic function $FK(\cdot)$ to transform the joint angles into their corresponding joint positions, calculating the joint loss $\mathcal{L}_\text{J} = \frac{1}{k}\sum_{j=1}^{k}\left\|FK\left(\boldsymbol{m}^j\right) - FK\left(\hat{\boldsymbol{m}}^j\right)\right\|_2^2$, where j represents the frame index and $\hat{\boldsymbol{m}}^j$ represents the predicted pose for this frame. We also compute velocity and acceleration, introducing the velocity loss: $\mathcal{L}_\text{V} = \frac{1}{k-1}\sum_{j=1}^{k-1}\left\|\left(\boldsymbol{m}^{j+1} - \boldsymbol{m}^j\right) - \left(\hat{\boldsymbol{m}}^{j+1} - \hat{\boldsymbol{m}}^j\right)\right\|_2^2$. Last, we apply the contact loss $\mathcal{L}_\text{C}$ which leverages binary foot-ground contact labels to optimize the consistency in foot contact during motion generation: $\mathcal{L}_\text{C} = \frac{1}{k-1}\sum_{j=1}^{k-1}\left\|\left(FK\left(\hat{\boldsymbol{m}}^{j+1}\right) - FK\left(\hat{\boldsymbol{m}}^j\right)\right)\cdot \hat{\boldsymbol{b}}^j\right\|_2^2$, where $\hat{\boldsymbol{b}}^j$ denotes the predicted binary foot-ground contact labels. The overall training loss is defined by the weighted sum of the above loss functions $\mathcal{L} = \mathcal{L}_\text{S} + \lambda_\text{J}\,\mathcal{L}_\text{J} + \lambda_\text{V}\,\mathcal{L}_\text{V} + \lambda_\text{C}\,\mathcal{L}_\text{C}$, where λ is are balancing parameters.

4 Experimental Setup and Results

Dataset: We evaluate the proposed method on FineDance dataset [35], which contains 7.7 hours of paired music and dance, totaling $831,600$ frames at 30 fps across different 16 genres. The average dance length is 152.3 seconds. The skeletal data of FineDance is stored in a 3D space and is represented by the standard 52 joints, including the finger joints. For all methods, we train on 183 pieces of music from the training set and generate 270 dance clips across 18 songs from the test set, using the corresponding real dances as ground truth.

Implementation Details: The training for our proposed approach uses two NVIDIA GeForce RTX 3090 GPU cards. The batch size and total number of epochs are set to 512 and 2000, respectively. The learning rate is set to $1e-4$, the hidden dimension is set to 512, and the guidance weight is set to 2.7. Four evaluation metrics are used in this paper, including Frechet inception distance (FID) [6], Diversity [6], Physical Foot Contact (PFC) [24] and Beat Alignment Score (BAS) [43].

Main Results: We first visualize the generated dance motions for three different music genres in Figure 2. The results are accompanied by the corresponding music, demonstrating that the proposed method is capable of generating realistic and complex dance movements with distinct characteristics. Then we compare the proposed method with four different music foundation models: CLAP [25], Wav2Vec [22], Jukebox [23] (used by the baseline model EDGE), and Wav2CLIP [19], as well as two sets of hand-crafted music features: STFT and 35-D Feature Set, which comprises 35-dimensional features provided by the FineDance dataset [35]. The results are reported in Table 1.

The experimental results indicate that the dance movements generated using music foundation models generally exhibit better physical realism and consistency with the music. Specifically, Wav2CLIP scores 0.171 in terms of PFC, while Jukebox achieves 0.223, attributed to their ability to capture a wide range of deep musical features. In contrast, the dances generated using hand-crafted features show better FID scores and diversity. For example, STFT produces FID scores of 18.497 for hands

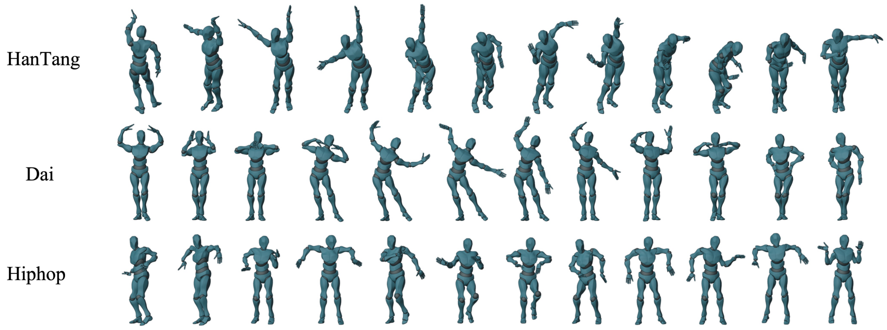

Figure 2: The generated dance motions for three different genres.

Table 1: A comparison of different music features.

	Motion Quality		Motion Diversity		PFC↓	BAS↑
	FID_hand↓	FID_body↓	Div_body↑	Div_hand ↑		
CLAP [25]	36.316	65.781	6.626	8.071	0.369	0.2256
Wav2Vec [22]	35.493	30.601	6.508	7.257	0.241	0.2226
Jukebox [23]	38.578	53.336	5.687	7.453	0.223	0.2161
Wav2CLIP [19]	27.212	30.539	6.683	7.396	**0.171**	**0.2283**
STFT	18.497	26.522	7.513	**8.562**	0.246	0.2201
35-D Feature Set [35]	21.681	27.092	7.458	8.022	0.262	0.2156
Wav2CLIP+STFT	**17.871**	**25.1752**	**7.758**	8.2543	0.209	0.2218

and 26.522 for the body, along with diversity scores of 7.513 and 8.562 for body and hands. These results suggest a closer match to the ground truth, although with diminished physical plausibility, as indicated by a PFC score of 0.241. The use of Wav2CLIP and STFT jointly produces the best results, combining the best FID score and diversity for the body, with other metrics also nearly approaching the best results. While music foundation models handle complex musical patterns well, this demonstrates that incorporating hand-crafted features improves generation of dance movements.

Table 2: A comparison with the state-of-the-art methods.

	Motion Quality		Motion Diversity		PFC↓	BAS↑
	FID_hand↓	FID_body↓	Div_body↑	Div_hand ↑		
DanceRevolution [44]	219.312	98.402	6.773	1.813	4.199	0.2171
Bailando [6]	45.083	52.373	4.943	5.629	0.361	0.2152
EDGE [24]	38.578	53.336	5.687	7.451	0.223	0.2161
DGFM	**20.417**	**23.648**	**7.631**	**8.10**	**0.207**	**0.2204**

We also compare our method with the state-of-the-art dance generation models, including DanceRevolution [44], Bailando [6], and EDGE [24]. As shown in Table 2, the proposed method achieves the best results across all the evaluation metrics. This highlights the advantages of combining features extracted from Wav2CLIP and STFT. The resulting movements not only exhibit high quality and diversity, but also maintain physical and rhythmic fidelity to the music.

5 Conclusion

In this paper, we have presented a new diffusion-based 3D dance generation approach using both music and text features. To investigate the impact of different foundation music models and hand-crafted music features on the motion generation task, we conducted extensive experiments. Specifically, we compared the generation results using four different foundation music models and two sets of hand-crafted music features. The experimental results demonstrated that the fusion of Wav2CLIP and STFT features achieves the best performance in terms of motion quality and synchronization with the music. Furthermore, we compared our model with several state-of-the-art models, and our model consistently outperformed them in multiple evaluation metrics.

References

[1] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "Ai choreographer: Music conditioned 3d dance generation with aist++," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 401–13 412.

[2] E. Aksan, M. Kaufmann, and O. Hilliges, "Structured prediction helps 3d human motion modelling," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7144–7153.

[3] H. Qiu, C. Wang, J. Wang, N. Wang, and W. Zeng, "Cross view fusion for 3d human pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4342–4351.

[4] A. Galata, N. Johnson, and D. Hogg, "Learning variable-length markov models of behavior," *Computer Vision and Image Understanding*, vol. 81, no. 3, pp. 398–413, 2001.

[5] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp, "Learn2dance: Learning statistical music-to-dance mappings for choreography synthesis," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 747–759, 2011.

[6] L. Siyao *et al.*, "Bailando: 3d dance generation by actor-critic gpt with choreographic memory," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 050–11 059.

[7] K. Chen *et al.*, "Choreomaster: Choreography-oriented music-driven dance synthesis," *ACM Transactions on Graphics (TOG)*, vol. 40, no. 4, pp. 1–13, 2021.

[8] B. Li, Y. Zhao, S. Zhelun, and L. Sheng, "Danceformer: Music conditioned 3d dance generation with parametric motion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 1272–1279.

[9] S. Yang, Z. Yang, and Z. Wang, "Longdancediff: Long-term dance generation with conditional diffusion model," *arXiv preprint arXiv:2308.11945*, 2023.

[10] Y. Huang *et al.*, "Genre-conditioned long-term 3d dance generation driven by music," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4858–4862.

[11] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE Access*, vol. 10, pp. 122 136–122 158, 2022.

[12] Q. Qi *et al.*, "Diffdance: Cascaded human motion diffusion model for dance generation," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 1374–1382.

[13] Y. Li *et al.*, "Mert: Acoustic music understanding model with large-scale self-supervised training," *arXiv preprint arXiv:2306.00107*, 2023.

[14] P.-Y. Huang *et al.*, "Mavil: Masked audio-video learners," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[15] X. Cheng, Z. Zhu, H. Li, Y. Li, and Y. Zou, "Ssvmr: Saliency-based self-training for video-music retrieval," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[16] D. McKee, J. Salamon, J. Sivic, and B. Russell, "Language-guided music recommendation for video via prompt analogies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 784–14 793.

[17] G. Mittal, J. Engel, C. Hawthorne, and I. Simon, "Symbolic music generation with diffusion models," *arXiv preprint arXiv:2103.16091*, 2021.

[18] J. Copet *et al.*, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[19] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, "Wav2clip: Learning robust audio representations from clip," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4563–4567.

[20] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[21] Y. Ma *et al.*, "Foundation models for music: A survey," *arXiv preprint arXiv:2408.14340*, 2024.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, “Jukebox: A generative model for music,” *arXiv preprint arXiv:2005.00341*, 2020.

[24] J. Tseng, R. Castellon, and K. Liu, “Edge: Editable dance generation from music,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 448–458.

[25] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[26] H. Liu *et al.*, “Audioldm: Text-to-audio generation with latent diffusion models,” *arXiv preprint arXiv:2301.12503*, 2023.

[27] D. Holden, J. Saito, and T. Komura, “A deep learning framework for character motion synthesis and editing,” *ACM Transactions on Graphics (TOG)*, vol. 35, no. 4, pp. 1–11, 2016.

[28] D. Holden, J. Saito, T. Komura, and T. Joyce, “Learning motion manifolds with convolutional autoencoders,” in *SIGGRAPH Asia 2015 technical briefs*, New York, NY, USA: Association for Computing Machinery, 2015, pp. 1–4.

[29] J. Butepage, M. J. Black, D. Kragic, and H. Kjellstrom, “Deep representation learning for human motion prediction and classification,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6158–6166.

[30] H. Chiu, E. Adeli, B. Wang, D.-A. Huang, and J. C. Niebles, “Action-agnostic human pose forecasting,” in *2019 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 2019, pp. 1423–1432.

[31] X. Du, R. Vasudevan, and M. Johnson-Roberson, “Bio-lstm: A biomechanically inspired recurrent neural network for 3-d pedestrian pose and gait prediction,” *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1501–1508, 2019.

[32] J. Li *et al.*, “Learning to generate diverse dance motions with transformer,” *arXiv preprint arXiv:2008.08171*, 2020.

[33] W. Zhuang, C. Wang, J. Chai, Y. Wang, M. Shao, and S. Xia, “Music2dance: Dancenet for music-driven dance generation,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 2, pp. 1–21, 2022.

[34] K. Gong *et al.*, “Tm2d: Bimodality driven 3d dance generation via music-text integration,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9942–9952.

[35] R. Li *et al.*, “Finedance: A fine-grained choreography dataset for 3d full body dance generation,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 234–10 243.

[36] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, “Vggsound: A large-scale audio-visual dataset,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 721–725.

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

[38] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to prompt for vision-language models,” *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[39] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

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

[41] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[42] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, “Human motion diffusion model,” in *The Eleventh International Conference on Learning Representations*.

[43] K. Onuma, C. Faloutsos, and J. K. Hodgins, “Fmdistance: A fast and effective distance function for motion capture data.,” *Eurographics (Short Papers)*, vol. 7, 2008.

[44] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang, “Dance revolution: Long-term dance generation with music via curriculum learning,” *arXiv preprint arXiv:2006.06119*, 2020.