# A Spatio-temporal Learning for Music Conditioned Dance Generation

Li Zhou
University of Massachusetts, Lowell
Lowell, USA

Yan Luo
University of Massachusetts, Lowell
Lowell, Massachusets, USA

## ABSTRACT

The music-conditioned dance generation, i.e., dancing to music, is a usage scenario of multi-modality human motion synthesis. Typically, it is a challenge to choreograph continuous motions coinciding with the melody and rhythm of the music. This paper proposes a position-wise encoding-decoding framework for spatio-temporal learning of motions and long-term skeleton-based dance generation oriented on music. Given the positional embedding of the frames in 1-minute video clips, firstly, we modularize a regional attention-based feed-forward mechanism to encode the music features. Secondly, based on the skeleton of each frame and the joint trajectories across motion frames, we formalize a graph topology to represent each dance sequence's spatial and temporal knowledge. Specifically, we propose a graph convolutional network (GCN) based blocks to process long-term dependencies of motions and leverage the spatial and temporal features. Both music and motion paths are learned fully in positional embedding schemes and constructed by repeating the corresponding blocks. Finally, as the task of dance generation is inherently the consistency between music and motions, we proposed a cross-modality feature fusion for multimodal interaction and music-conditioned dance generation. Experimental results demonstrate that our method outperforms state-of-art methods in motion quality and motion-music correlation metrics.

## CCS CONCEPTS

• **Computing methodologies** → **Artificial intelligence**.

## KEYWORDS

Attention, Graph Topology, Graph Neural Networks

**ACM Reference Format:**
Li Zhou and Yan Luo. 2022. A Spatio-temporal Learning for Music Conditioned Dance Generation. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3536221.3556618

## 1 INTRODUCTION

Dancing to music is a natural cross-modal behavior of humans that matches movement patterns with musical beats. The choreography is the harmonious combination of music and motion data [7, 8]. In terms of deep learning, the music-conditioned dance generation is a creative process of multimodal interaction. A dancing pose at a timestamp can synthesize the next in various modes, resulting in a long-range dance generation that leads to high kinematic complexity. Besides, the dance generation necessitates the synchronization of motions with the music in melody and rhythm. Therefore, the sequence of composing dance faces the challenges of both motion continuousness and music consistency [17, 22]. This paper addresses the challenges by proposing a two-path spatio-temporal learning framework for long-term dance generation synchronized with music features. To ensure the alignment of the motion and audio sequences, we apply a positional encoding mask for the series throughout the framework. In other words, the paired motion-music data is parallel computed.

The framework mainly consists of three parts, as illustrated in Fig. 1: audio encoding, motion decoding, and multimodal interaction. We introduce a regional attention-based feed-forward mechanism to encode the sparsity music features. Attention mechanism [31] is one of the robust strategies to represent the content of inputs in natural language processing. Existing methods [25, 29, 34, 36] validate the efficiency of attention for long sequence information extraction. Inspired by the successful use of the attention mechanism in the sequence data, we introduce a new attention-based layer to encode the sparse representation of music features. The layer models sparsity features into high-level learned features of each timestamp. To improve the robustness of the attention mechanism for long-term sequence, we mask out the subsequent information and control the perceptive field in the attention with a sliding window technique.

Furthermore, we formalize the graph topology of each sequence for motion feature decoding. The graph data is essential to the decoding path as it models the spatial information of a pose in a frame and the temporal adjacency across the sequence. Regarding the body joints as vertexes, we draw the spatial edges depending on the natural connections of joints in human bodies; we draw the temporal edges according to the same joints following the timestamps. Graph neural networks (GNNs) are deep learning algorithms built explicitly for non-Euclidean graphs [30, 35]. Recent works show the ability of GNNs to learn the information of the skeleton-based graph structure for the tasks of action recognition [5, 10, 18, 21] and motion prediction [1, 23, 33]. Inspired by the previous works, we propose a GCN-based spatio-temporal block to decode the motion features and learn the long dependencies of motions. The spatial convolutional part interprets information from the channels of the joints at each timestamp. And the temporal convolutional part refines the dependencies from the time series of each joint.


*ICMI '22, November 7–11, 2022, Bengaluru, India*

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556618

Finally, we propose an attention-based cross-modality mechanism to fuse the features between music and motions and generate the music-conditioned dance. The attention mechanism to consecutive data alleviates error accumulation compared with RNNs [1, 27, 31]. The multimodal attention mechanism is powerful for feature interactions to depict guided interactions and interpret the multimodal features [6, 36]. We exploit the proposed mechanism's weights to transform the music-motion sequence that led to the music-to-dance generation.

Our contributions can be summarized as follows. (1) We construct the music conditioned dance generation as a long-term sequence-to-sequence multimodal task and introduce a novel spatio-temporal learning framework for skeleton-based dance generation. (2) We formalize a graph topology to model the skeleton and joint trajectories and propose a positional GCN-based block to decode spatial and temporal features of motions. (3) We introduce a regional attention-based encoding mechanism for self feature learning and mutual feature fusion. (4) Experimental results show that our framework outperforms state-of-art methods concerning motion quality and motion-music consistency.

## 2 RELATED WORK

Studies on the prediction of motion sequences are well exploited using recurrent neural networks (RNNs) [9, 11, 24, 27] because capable of learning temporal dependencies. However, RNN based methods often produce unrealistic predictions [24, 27], i.e. freeze poses, because using the inference pose as input to the next estimation often result in error accumulation throughout the prediction sequence. Also, the models [9, 15] can easily fail into discontinuousness between the last inference pose and the first one. Because of the limitations of RNNs, other studies use non-recurrent models as an alternative for the motion prediction. For example, Li et al. [19] improve the long-term dependencies by CNNs to predict the human motion. Butepage et al. [3] propose a fully convolutional network with a feed-forward temporal encoder to exploit the pose history. Besides, GCNs are used to study human motions by encoding the structure of human pose into graph topologies [5, 18, 26]. Mao et al. [26] encode the short-term history of motions via discrete cosine transform (DCT) and train a GCN to learn spatio-temporal dependencies for motion prediction.

The audio to human motion generation is typical cross-modal learning task that generate music conditioned motions. Based on the motion prediction methods, studies on the task focus on the multimodal composition and synchronization. Zhuang et al. [37] model a conditional distribution with an autoregressive generative model to generate music conditioned dance. Li et al. [20] utilize transformer based model on motion and audio respectively, and then compose features to generate diverse dance. Similarly, Huang et al. [14] propose a sequence-to-sequence model with a concatenation operation to fuse fine-grained features of motion and audio.

## 3 METHODOLOGY

This section elaborates on the two-path framework for music conditioned skeleton-based dance generation. Following the overview illustrated in Fig. 1, spatio-temporal learning of the task is summed up in two parts shown in Fig. 2: graph-based decoding and attention-based encoding. Formally, the input dataset $D = \{(X_i, Y_i)\}_{i=1}^{\mathbb{N}}$ consists of $\mathbb{N}$ paired motion-music sequences. An audio sequence sample is denoted as $X = \{x_i\}_{i=1}^{T}$, where $x_i \in \mathbb{R}^{d_x}$ is a vector of audio features at timestamp $i$ and the length of the sequence is $T$. A motion sequence sample is denoted as $Y = \{y_i\}_{i=1}^{T}$, where $y_i = \{v_{ij}\}_{j=1}^{N}$ is a pose representation of $N$ joints in a frame at timestamp $i$, and $v_{ij} \in \mathbb{R}^{C}$ is a $j$-th joint of a pose being a vector of a joint's channels at timestamp $i$. Intuitively, $x_i$ and $y_i$ are synchronized at each timestamp. The problem is to build a generation model $g : X \rightarrow Y$ that estimates a sequence of new dance $Y$ that oriented on music style.

### 3.1 Attention-based Encoding

Inspired by the appealing performance on attention mechanism [31] on sequence data for self-interactions [13, 25] and mutual-interactions [1, 36], we introduce a regional attention-based encoding layer as shown in Fig. 2 (a), and stack $L_x$ identical encoding layers to refine the input sequence. Specifically, we extract audio features of a piece of music as an input sequence for audio feature learning in the audio path and concatenate learned features from the two-path processing as an input sequence for audio-to-motion feature composition.

We first project an input sequence $X$ into a source input embedding via a single linear layer ($\mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_{x'}}$). Unlike RNNs for CNNs, attention has no concept of order. So, we also embed the sequence $X$ into a position encode embedding with a sinusoidal position encoding table ($\mathbb{R}^{d_x} \rightarrow \mathbb{R}^{d_{x'}}$). Thus, for a sequence of input $X = \{x_i\}_{i=1}^{T} \in \mathbb{R}^{T \times d_x}$, the embedding output is the positional add of the above mentioned two embeddings and results in $\tilde{X} = \{\tilde{x}_i\}_{i=1}^{T} \in \mathbb{R}^{T \times d_{x'}}$.

Then, the embeddings are learned by the regional multi-head attention. Recall the scaled dot-product attention, proposed by [31], be operated in the quadratic term $O(T^2)$ in both space and time complexities and costs in huge memory for long sequences with length $T$. We mask out the subsequent information and control the attention field for query and key-value representations by a sliding window with size $n$ ($n <= T$) and reduces the space complexity into $O(Tn)$. The window size can be small if we address on the local pattern of a sequence, such as a clip of music representation. The attention operation is formulated as:

$$\begin{aligned} Q = \tilde{X}W^{Q}, K = \tilde{X}W^{K}, V = \tilde{X}W^{V} \\ Attention(Q, K, V, M) = softmax(\frac{QK^{\top}}{\sqrt{d_k}} + M)V \end{aligned} \tag{1}$$

where $d_k$ is the kernel size, mask $M \in \mathbb{R}^{d_k \times d_k}$, query, key and value $Q, K, V \in \mathbb{R}^{T \times d_k}$ leverage from weight matrices $W^{Q}, W^{K}, W^{V} \in \mathbb{R}^{d_{x'} \times d_k}$ respectively. Multi-head attention (MHA) [31] employs $h$ scaled dot-product attentions (referred as 'heads') and concatenates learned features $H_i$ using the weighted function. Finally, the attention output $X^{'}$ can be formulated as:

$$\begin{aligned} H_i = Attention(Q_i, K_i, V_i, M) \\ X^{'} = \boldsymbol{f}(Concat(H_1, ..., H_h)W^{o}) \end{aligned} \tag{2}$$

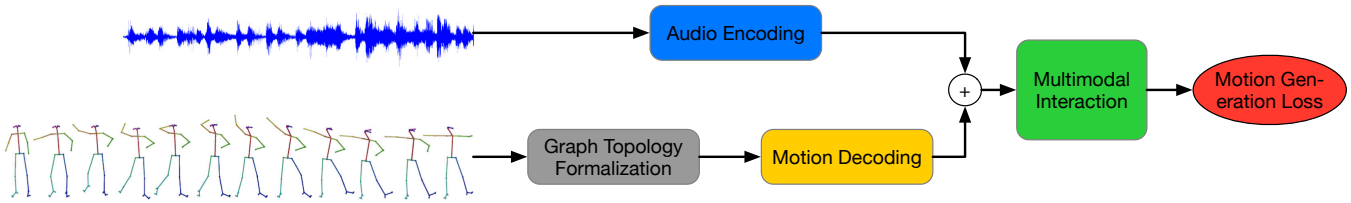
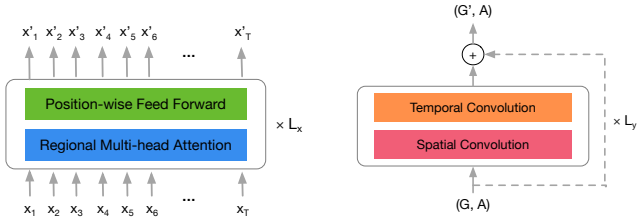



**Figure 1: Cross-Modal Dance Generation Overview. We perform dance generation in a two-path spatio-temporal learning mode synchronized with audio features. The audio path consists of blocks of audio encoding, and the motion path consists of spatial-temporal graph topology formalization and layers of motion decoding. The output is a sequence of music-conditioned motions generated from multimodal interaction.**




(a) attention-based encoding layer (b) graph-based decoding layer

**Figure 2: Details of Layers (a) Attention-based encoding layer encodes a sequence $X = (x_1, ..., x_T)$ into a hidden matrix $X' = (x'_1, ..., x'_T)$; (b) Graph-based decoding layer takes an undirected graph represented feature map $G$ and an adjacency matrix $A$ as inputs, and generates learned features $G'$.**

where $i$-th head of query, key and value $Q_i, K_i, V_i \in \mathbb{R}^{T\times d_{k'}}$, and $d_{k'} = d_k/h$, weight matrix $W^o \in \mathbb{R}^{T\times d_k}$, output $X^{'} \in \mathbb{R}^{T\times d_{x'}}$, and $f(\cdot)$ is a position-wise feed forward composed of two feed-forward networks (FFNs). FFN is a sequence of a $1D$ convolution, a ReLU, a dropout, a residual connection and a layer normalization.

## 3.2 Graph-based Decoding

To decode the spatio-temporal structure of motions in the motion path, as shown in Fig. 1, is composed of two sections. The first section, named spatio-temporal graph topology formalization, depicts the intra- and inter-skeleton connections in a sequence of motions. The second section is called motion decoding, which stacks graph-based decoding layers as illustrated in Fig. 2 (b).

*3.2.1 Graph Topology Formalization.* Given a sequence of frames $Y = \{y_i\}_{i=1}^{T} \in \mathbb{R}^{T\times d_y}$ and a set of joints $N$ per frame, we focus on learning the co-occurrence pattern of intra- and inter-skeleton features. In this section, we represent a skeleton sequence hierarchically as an undirected graph to feature joints' spatial and temporal connections. The spatial connections are defined by the natural connections of joints in human bodies shown in Fig. 3 (a). The temporal connections are trajectories of the same joints following the timestamps displayed in Fig. 3 (b).

Formally, the undirected graph $G = (V, E)$ is composed from $T$ frames and $N$ joints per frame. The node matrix $V = \{v_{ij}\}_{i=1,j=1}^{T,N} \in \mathbb{R}^{T\times N\times C}$ includes all joints in a skeleton sequence, and $v_{ij} \in \mathbb{R}^C$ indicates the $j$-th node at timestamp $i$ has a feature vector in $C$ channels. The edge set $E = \{E_S, E_T\}$ is made up of spatial and temporal connections. $E_S = \{(v_{ij}, v_{ij'})\}$ denotes the spatial edges, including the skeleton connections as described in Fig. 3 (a) and self connections of each joint ($j = j'$). And $E_T = \{(v_{ij}, v_{(i+1)j})\}$ denotes the temporal edges.

The adjacency matrix A is the key factor for the graph topology learning. As the temporal edges are well-ordered, we simplify the problem by focusing on the representation of the spatial edges in a single frame. We first get the uniform adjacency matrix $H = \{h_{ij}\} \in \mathbb{R}^{N\times N}$ by the $l$-hop distance method in a frame, i.e. $h_{ij}$ is a connectable distance from the node $v_i$ to $v_j$ within $l$ steps, including the self connections. Then, we get a normalized uniform matrix $H'$ by $\Lambda^{-1/2}H\Lambda^{-1/2}$ [16], where $\Lambda = \{\sum_{j=1}^{N} h_{ij}\}$. Finally, we build the adjacency matrix $A$ based on the matrix $H'$.

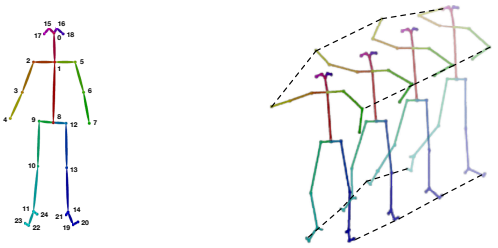
Inspired by the partition strategies discussed in [32], we use the spatial partition strategy to compose $A$ in multiple scales of connections based on the uniform matrix $H$, that is, a node is labeled by the property of the distance to a central node $v_c$. In this work, the adjacency matrix is defined in three levels of connections $A = \{\{a_{0ij}\}, \{a_{1ij}\}, \{a_{2ij}\}\} \in \mathbb{R}^{3\times N\times N}$, and the connections are defined by:

$$a_{0ij} = \begin{cases} h'_{ij}, & \text{if } h_{ic} = h_{jc} \\ 0, & \text{otherwise} \end{cases}, a_{1ij} = \begin{cases} h'_{ij}, & \text{if } h_{ic} < h_{jc} \\ 0, & \text{otherwise} \end{cases},$$
$$a_{2ij} = \begin{cases} h'_{ij}, & \text{if } h_{ic} > h_{jc} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $h_{ic}$ is a distance between node $v_i$ to the central node $v_c$.

*3.2.2 Spatio-temporal Convolution.* As shown in Fig. 2 (b), the motion decoding is formed by stacking $L_y$ graph-based decoding layers together, which depicts the input sequence of motions in a graph topology. The graph-based decoding layer is the sequential spatial and temporal convolution processing. The residual connection accompanied by the decoding layer maps an identity topology with spatio-temporal convolutional outputs that address the performance degradation in deep neural architectures.

The input of the layer is a set of $(G, A)$ generated from the graph topology formalization section, where $G \in \mathbb{R}^{C\times T\times N}$ and $A \in \mathbb{R}^{C\times N\times N}$. The spatial convolution is a module for a graph convolution of each frame, which consists of a $2D$ convolution, a matrix multiplication between graph and adjacency matrix. The temporal convolution is a module for temporal connections with a



(a) skeleton layout (b) the spatio-temporal graph

**Figure 3: Graph Topology of Skeleton (a) skeleton layout of spatial connections between human joints, the number of joints per frame is $N = 25$; (b) the spatio-temporal graph of a motion sequence, the dash lines are examples of temporal connections over consecutive timestamps.**

sequence of a 2$D$ batch normalization (BN), a ReLU, a 2$D$ convolution, a 2$D$ BN and a dropout. The output is a decoded graph $G'$ in terms of channels. The layer can be formulated as:

$$\begin{aligned} Spatial(G, A) &= GW^E \otimes A \\ G' &= Temporal(Spatial(G, A)) + G \end{aligned} \tag{4}$$

where $W^E \in \mathbb{R}^{(d_k \times C) \times C \times d_g \times d_g}$ is the trainable weight matrix and $d_g$ is the graph kernel size, $\otimes$ is an element-wise product and sums out the dimension of channels $C$, the output $G' \in \mathbb{R}^{d_k \times T \times N}$. Particularly, the decoded graph $G'$ from the last layer in the motion decoding is followed by a linear function to get the refined motion feature $Y' \in \mathbb{R}^{T \times d_{y'}}$.

Finally, As mentioned in section 3.1, we use the regional attention mechanism for audio-motion feature composition. The audio features $X'$ and motion features $Y'$ are concatenated along the first dimension $T$. Followed by the multimodal interaction, the output $Z \in \mathbb{R}^{T \times d_y}$ is generated by comparing with the ground truth motions $Y$ with minimized $L1$ loss.

# 4 EXPERIMENTS

## 4.1 Experimental Setup

*4.1.1 Dataset.* All of the experiments in this study were performed using the dataset collected by [17]. The dataset contains approximate 71 hours video clips including 3 styles of dance: "Ballet", "Hiphop" and "Zumba". We construct the dataset by the methods mentioned in [14]. To fit the long-term music-conditioned dance generation task on the position-wise framework, we prune and extract the dataset into one-minute clips of audio data and motion data in 15 frames per second (FPS). Thus, we acquire 136 clips for "Ballet", 298 clips for "Hiphop" and 356 clips for "Zumba". We do a stratified split on the dataset in a 90% − 10% fashion for training and test sets.

For audio feature generation, we use Librosa [28] with sampling rate at 15400 Hz to get MFCC and MFCC delta, and decompose audio sequences into harmonic and percussive components to generate tempogram, chromagram and onset strength. We use a dynamic programming beat tracker [28] to detected the beat information from onset strength, and represent results into a one-hot vector. We concatenate extracted features into the final audio feature with $d_x = 438$. For skeleton-based motion generation, we use OpenPose [4] to extract 2$D$ body key joints from each clip. As shown in 3 (a), each frame is made up of 25 key joints and results in a vector with $d_y = 50$.

*4.1.2 Implementation Details.* Following the dance generation framework in Fig. 1, the input contains one-minute sequences with 15 FPS, i.e. $T = 900$ for both a motion sequence and an audio sequence, where the sequences are synchronized on each frame. The hyper-parameters of the framework are listed as follows: both audio encoding and multimodal interaction modules consist of $L_x = 1$ attention-based encoding layer with $h = 8$ heads, $d_k = 64$ and 1024 hidden units. The motion decoding module consists of $L_y = 4$ identical graph-based decoding layers, with 15 temporal kernel size, 3 spatial kernel size, 1024 hidden units and 64 out channels. We train the model with 16 batch size using Adam optimizer with the base learning rate 1e-4 on 2 NVIDIA Tesla V100.

## 4.2 Evaluation Metrics

We evaluate the motion quality of generated dances by the Average Spatial Distance and the Frechet Distance. The average spatial distance is defined through the average Euclidean distance between a cluster of points in a 2$D$ plane, which is used to measure the similarity between skeletons [2]. The closer the spatial distance to the real dance, the better the realism of generated motions is. Frechet Inception Distance (FID) [12] evaluate the distribution distance between a generated sequence and a ground-truth sequence. There is no standard inception networks for motion evaluation. In this work, we measure the spatio-temporal motion quality by calculating the Frechet Distance (FD) directly from the synthesized joint positions $Y'$ and the ground-truth $Y$ (lower is better).

Furthermore, we evaluate the dance style consistency by the Beat Alignment Score introduced in [22], i.e. score the correlation between motion beats and audio beats. The motion beats are calculated as the relative minima in kinematic velocity. The audio beats are the 1-dim one-hot beats, which are generated from Librosa [28]. The Beat Alignment Score is defined as the average distance of each motion beat that are aligned to its nearest audio beat:

$$score = \frac{1}{n} \sum_{k=1}^{m} exp(min(\|bx_i, by_j\|^2)/(2\sigma^2)) \tag{5}$$

where $Bx = \{bx_i\}$ and $By = \{by_j\}$ is the audio beats and motion beats respectively, $\sigma = 3$ is a normalize factor.

## 4.3 Quantitative Evaluation

In this section, we report the experimental results of the encoding-decoding framework with the two baselines: [17] and [14] on the test set as mentioned in section 4.1.1. The results are shown in Table 1. Compared with the two baselines, our generated motion sequences overall are much closer to the real dances, as well as three styles of dances, in the aspect of spatio-temporal evaluation by FD. While Dance Revolution produces closer FD in Zumba, the average spatial distance difference between the real dances and Dance Revolution is worse than between real and our methods. Generated motions from Dancing2Music perform better than our method regarding the spatial distance only but worse in the temporal part

**Table 1: Dance Generation Evaluation**

| | Ballet | | | Hiphop | | |
|---|---|---|---|---|---|---|
| Methods | Spatial Dist. | Frechet Dist.↓ | Beat Align. | Spatial Dist. | Frechet Dist.↓ | Beat Align. |
| Real Dances | 9.406 | - | 0.382 | 9.476 | - | 0.389 |
| Dancing2Music [17] | 7.211 | 64.413 | 0.371 | 6.925 | 49.717 | 0.389 |
| Dance Revolution [14] | 5.622 | 66.752 | 0.373 | 4.568 | 48.505 | 0.397 |
| **Ours** | 5.981 | 54.381 | 0.381 | 5.598 | 44.619 | 0.398 |

| | Zumba | | | Overall | | |
|---|---|---|---|---|---|---|
| Methods | Spatial Dist. | Frechet Dist.↓ | Beat Align. | Spatial Dist. | Frechet Dist.↓ | Beat Align. |
| Real Dances | 9.382 | - | 0.420 | 9.421 | - | 0.397 |
| Dancing2Music [17] | 9.772 | 94.308 | 0.420 | 7.969 | 69.479 | 0.393 |
| Dance Revolution [14] | 6.036 | 54.259 | 0.419 | 5.409 | 56.505 | 0.397 |
| **Ours** | 7.834 | 61.658 | 0.418 | 6.471 | 53.553 | 0.398 |

evaluation, which indicates the motions lack consistency. Overall, our generated motion sequences is more realistic in the combination of spatial and temporal evaluation Besides, we evaluate the dance style consistency with music by the beat alignment score. Overall, our method and Dance Revolution result in scores very close to the real dances. However, we observe that the alignment score is better correlated with Ballet's real dances than the two baselines. This shows that our method for motion-music correlation is better than the others in general.

## 5 CONCLUSION

This paper presents a novel position-wise encoding-decoding framework for spatio-temporal learning of long-term skeleton-based motions and music conditioned motion generation. Specifically, we introduce a regional attention-based encoding layer to efficiently learn long-term sequences of audio features and fuse multimodal sequences of audio-to-motion features. Besides, we propose a graph topology formalization method to depict the intra- and inter-skeleton connections in a motion sequence and present a graph-based decking layer to interpret the spatio-temporal information from the graph. The experimental results show a promising performance of our framework on motion quality and motion-music correlation metrics.

## REFERENCES


[1] Emre Aksan, Manuel Kaufmann, Peng Cao, and Otmar Hilliges. 2021. A spatio-temporal transformer for 3D human motion prediction. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 565–574.

[2] Vladislav Ayzenberg and Stella F Lourenco. 2019. Skeletal descriptions of shape provide unique perceptual information for object recognition. *Scientific reports* 9, 1 (2019), 1–13.

[3] Judith Butepage, Michael J Black, Danica Kragic, and Hedvig Kjellstrom. 2017. Deep representation learning for human motion prediction and classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6158–6166.

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

[5] Li Chaolong, Cui Zhen, Zheng Wenming, Xu Chunyan, and Yang Jian. 2018. Spatio-temporal graph convolution for skeleton based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[6] Shaoxiang Chen and Yu-Gang Jiang. 2019. Motion guided spatial attention for video captioning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 8191–8198.

[7] Lucie Clements, Emma Redding, Naomi Lefebvre Sell, and Jon May. 2018. Expertise in evaluating choreographic creativity: An online variation of the consensual assessment technique. *Frontiers in psychology* (2018), 1448.

[8] Dieter Drobny and Jan Borchers. 2010. Learning basic dance choreographies with different augmented feedback modalities. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. 3793–3798.

[9] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. 2015. Recurrent network models for human dynamics. In *Proceedings of the IEEE international conference on computer vision*. 4346–4354.

[10] Xiang Gao, Wei Hu, Jiaxiang Tang, Jiaying Liu, and Zongming Guo. 2019. Optimized skeleton-based action recognition via sparsified graph regression. In *Proceedings of the 27th ACM International Conference on Multimedia*. 601–610.

[11] Anand Gopalakrishnan, Ankur Mali, Dan Kifer, Lee Giles, and Alexander G Ororbia. 2019. A neural temporal model for human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12116–12125.

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* 30 (2017).

[13] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music transformer. *arXiv preprint arXiv:1809.04281* (2018).

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

[15] Ashesh Jain, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-rnn: Deep learning on spatio-temporal graphs. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 5308–5317.

[16] Stefanie Jegelka. 2022. Theory of Graph Neural Networks: Representation and Learning. *arXiv preprint arXiv:2204.07697* (2022).

[17] Hsin-Ying Lee, Xiaodong Yang, Ming-Yu Liu, Ting-Chun Wang, Yu-Ding Lu, Ming-Hsuan Yang, and Jan Kautz. 2019. Dancing to music. *Advances in Neural Information Processing Systems* 32 (2019).

[18] Bin Li, Xi Li, Zhongfei Zhang, and Fei Wu. 2019. Spatio-temporal graph routing for skeleton-based action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8561–8568.

[19] Chen Li, Zhen Zhang, Wee Sun Lee, and Gim Hee Lee. 2018. Convolutional sequence to sequence model for human dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5226–5234.

[20] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. 2020. Learning to generate diverse dance motions with transformer. *arXiv preprint arXiv:2008.08171* (2020).

[21] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3595–3603.

[22] Ruilong Li, Shan Yang, David A Ross, and Angjoo Kanazawa. 2021. Learn to dance with aist++: Music conditioned 3d dance generation. *arXiv e-prints* (2021), arXiv–2101.

[23] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. 2019. Grip: Graph-based interaction-aware trajectory prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 3960–3966.

[24] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2017. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363* (2017).

[25] Wei Mao, Miaomiao Liu, and Mathieu Salzmann. 2020. History repeats itself: Human motion prediction via motion attention. In *European Conference on Computer Vision*. Springer, 474–489.

[26] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. 2019. Learning trajectory dependencies for human motion prediction. In *Proceedings of the*

*IEEE/CVF International Conference on Computer Vision*. 9489–9497.

[27] Julieta Martinez, Michael J Black, and Javier Romero. 2017. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2891–2900.

[28] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, Vol. 8. Citeseer, 18–25.

[29] Hai-Feng Sang, Zi-Zhen Chen, and Da-Kuo He. 2020. Human motion prediction based on attention mechanism. *Multimedia Tools and Applications* 79, 9 (2020), 5529–5544.

[30] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks* 20, 1 (2008), 61–80.

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

[32] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

[33] Wenjie Yin, Hang Yin, Danica Kragic, and Mårten Björkman. 2021. Graph-based Normalizing Flow for Human Motion Generation and Reconstruction. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 641–648.

[34] Xiang Zhang, Lina Yao, Salil S Kanhere, Yunhao Liu, Tao Gu, and Kaixuan Chen. 2018. Mindid: Person identification from brain waves through attention-based recurrent neural network. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–23.

[35] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81.

[36] Li Zhou and Yan Luo. 2021. Deep Features Fusion with Mutual Attention Transformer for Skin Lesion Diagnosis. In *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 3797–3801.

[37] Wenlin Zhuang, Congyi Wang, Jinxiang Chai, Yangang Wang, Ming Shao, and Siyu Xia. 2022. Music2Dance: DanceNet for Music-Driven Dance Generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 2 (2022), 1–21.