DanceMVP: Self-Supervised Learning for Multi-Task Primitive-Based Dance Performance Assessment via Transformer Text Prompting

Yun Zhong, Yiannis Demiris*

Personal Robotics Lab, Dept. of Electrical and Electronic Engineering,
Imperial College London
{y.zhong20,y.demiris}@imperial.ac.uk

Abstract

Dance is generally considered to be complex for most people as it requires coordination of numerous body motions and accurate responses to the musical content and rhythm. Studies on automatic dance performance assessment could help people improve their sensorimotor skills and promote research in many fields, including human motion analysis and motion generation. Recent papers on dance performance assessment usually evaluate simple dance motions with a single task - estimating final performance scores. In this paper, we propose **DanceMVP**: **m**ulti-task dance performance assessment **v**ia text **p**rompting that solves three related tasks - (i) dance vocabulary recognition, (ii) dance performance scoring and (iii) dance rhythm evaluation. In the pre-training phase, we contrastively learn the primitive-based features of complex dance motion and music using the InfoNCE loss. For the downstream task, we propose a transformer-based text prompter to perform multi-task evaluations for the three proposed assessment tasks. Also, we build a multimodal dance-music dataset named **ImperialDance**. The novelty of our ImperialDance is that it contains dance motions for diverse expertise levels and a significant amount of repeating dance sequences for the same choreography to keep track of the dance performance progression. Qualitative results show that our pre-trained feature representation could cluster dance pieces for different dance genres, choreographies, expertise levels and primitives, which generalizes well on both ours and other dance-music datasets. The downstream experiments demonstrate the robustness and improvement of our method over several ablations and baselines across all three tasks, as well as monitoring the users' dance level progression.

Introduction

Dance is a form of art with highly complex multi-sensory processes. Experienced dancers could perform numerous movements and align these movements to musical beats, which requires professional practice and training to equip them with such abilities. However, professional training is usually expensive and limited by time and geography. Hence, a multi-task dance performance assessment model that continuously evaluates a dancer's performance quality could be developed to address this challenge.

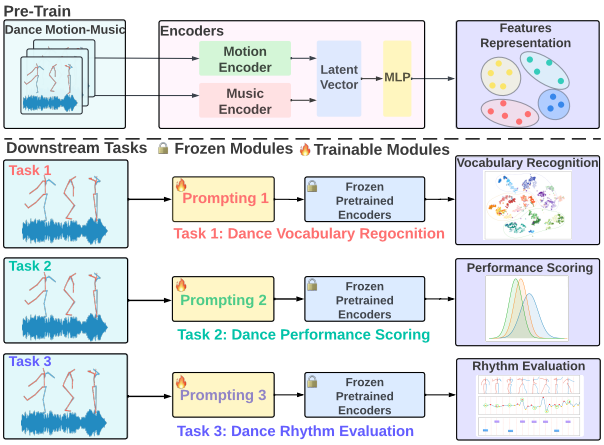

Figure 1: Proposed dance performance assessment model. The input is dance motion and music, the output is the multi-task performance evaluation including (i) dance vocabulary recognition: dance genres, choreographies and expertise levels classification for recognized motion primitives, (ii) dance performance scoring and (iii) dance rhythm evaluation: musical beats and motion alignment rate regression.

Previous studies on Action Quality Assessment (AQA) and Skill Determinations/Assessment (SA) have been mainly focusing on surgery (Liu et al. 2021), Olympics sports (Parmar and Morris 2019; Xu et al. 2022) and human daily activities (Doughty, Mayol-Cuevas, and Damen 2019) instead of dancing. Some recent research on multimedia dance-music tasks targeting dance motion generation (Li et al. 2022; Siyao et al. 2022), motion and music synthesis (Ren et al. 2020a; Chen et al. 2021), rather than dance-music performance assessment. Other dance motion analysis approaches (Chan et al. 2011a; Zhong, Zhang, and Demiris 2023) have focused on simple, controlled conditions, where their models ignore the music features or only provide final performance scores. Hence, two main challenges of automatic dance performance assessment could be summarized: (i) feature learning of complex dance motions and music: The feature representations of various styles, genres, and

*Yiannis Demiris is supported by a Royal Academy of Engineering Chair in Emerging Technologies.
Code&Dataset at www.imperial.ac.uk/personal-robotics/software

choreographies of dancing are different. The feature representation of the same dance sequence performed by people with different expertise levels also varies; (ii) characterizing the multi-task dance performance assessment process. Evaluating dance performance is a multi-task process that contains an interpretation of various multimedia content, including identifying the elements and types of the performed dance sequences, determining the score of each element, and measuring the relations between motions and music.

Our method aims to tackle the problems mentioned above related to dance performance assessment, especially in complex, real-world professional dance training scenarios. Fig. 2 illustrates the self-supervised workflow of our study. We introduce a contrastive learning framework to learn the feature representation of the primitive-based dance sequences performed by people with different dance expertise levels. The dance sequences include diverse dance music, genres and choreographies. Instead of assessing the performance based on the entire input dance sequences (Zhong, Zhang, and Demiris 2023), we will be training the model with music and motion primitives. Each motion and music sequence is divided into primitive pieces by using the Eight-Beats segmentation method, which characterizes the rhythmic structure in motion and music data (Royal Academy of Dance 2020). We then deploy Spatial-Temporal Graph Convolutional Network (ST-GCN) (Yan, Xiong, and Lin 2018) and Deep Recurrent Network (LSTM) (Hochreiter and Schmidhuber 1997) encoders to learn the dance primitives' features.

Having the feature representations of the motion-music primitives, we then deploy a transformer-based text prompting method (Radford et al. 2021; Brown et al. 2020; Jia et al. 2022) to evaluate the dance performance in three downstream tasks. For our formulated dance performance assessment problem, numerous diverse dance genres and choreographies are emerging worldwide. Training a new model with new incoming data frequently is not always feasible, and to enable a dance model to provide accurate predictions, it usually requires a large number of samples of various dance sequences for the same condition (e.g., same choreography, same music, or at the same time step), which is often challenging to collect and label such sufficient samples. Hence, we deploy prompt tuning to our pre-trained model to perform downstream tasks with complex and unseen dance motions and music to solve the above-mentioned challenges. Experiments show that our prompt tuning outperforms the fine-tuning in three downstream dance performance assessment tasks: (i) dance vocabulary recognition; (ii) dance performance scoring, and (iii) dance rhythm evaluation.

We also construct the ImperialDance dataset, which comprises 69,300 seconds of dance motions, five genres, 20 pieces of music and 20 choreographies with dancers of 3 different expertise levels. Quantitative and qualitative experiments have been conducted to demonstrate our approach's effectiveness in three evaluation tasks and level progression monitoring. Our contributions are summarized as follows:

- We propose a contrastive self-supervised learning framework to represent the features of primitive-based motion and music pieces. These pieces belong to different dance genres, choreographies and expertise levels.
- We propose a transformer text prompter, artistically characterizing the downstream multi-tasks dance performance assessment process, including (i) dance vocabulary recognition: recognition of dance motion primitives, genres, choreographies, and expertise levels; (ii) dance performance scoring: quality score distribution estimation and (iii) dance rhythm evaluation: music-motion alignment rate regression.
- We construct the ImperialDance dataset, the first dataset containing dance motions for different expertise levels and a significant amount of repeating dance sequences for the same choreography and music, making the collected data especially useful for dancer's expertise level monitoring and music-dance synthesis evaluation.

Related Works

Human Motion Analysis Analyzing human motion for motion generation (Li et al. 2022), rehabilitation (Sjödahl Hammarlund et al. 2002) and sports competition (Bertasius et al. 2017) has been actively researched recently. As for a particular scenario - dancing analysis (i.e., automatic and quantitative analysis of how well a dance motion is performed) has received increasing attention from researchers, thanks to the development of motion analysis and multimedia signal processing (Alexiadis and Daras 2014; Chan et al. 2011b). However, these methods focus on evaluating simple dance motions with specific dance choreography and expertise levels, making them hard to generalise to real-world assistive technologies. Zhong et al. 2023 has introduced a self-supervised method to assess dancers' expertise levels based on different dance genres and choreographies. The other assessment tasks, such as dance rhythm and performance score evaluation, still need to be investigated.

Action Quality Assessment (AQA) Besides motion analysis, determining the skill level of human motion is also related to the dance performance assessment. The majority of prior works are focusing on surgery (Sharma et al. 2014) or Olympic sports level determination (Tang et al. 2020) instead of dancing. These works demonstrate good performance by focusing on representing the features of surgery or Olympic tasks. However, there is less research in the literature concentrating on dance quality assessment. The approaches that focus on surgery or sports are generally challenging to apply to dancing, as they tend to only explore the motion feature representation, ignoring the feature correlation between motion and other elements, like music rhythm.

Self-Supervised Learning Self-supervised learning has made progress across different applications (Sumer, Dencker, and Ommer 2017; Ren et al. 2020b; Zhang and Demiris 2022). As for human motion analysis, self-supervised learning has also gained popularity because of its ability to learn feature representations from data without introducing label bias. Recent self-supervised learning works have focused on learning the similarity between sample pairs by utilizing contrastive loss. Study (Chen et al. 2020) shows that contrastive self-supervised learning has enabled the model to achieve state-of-the-art performance in image

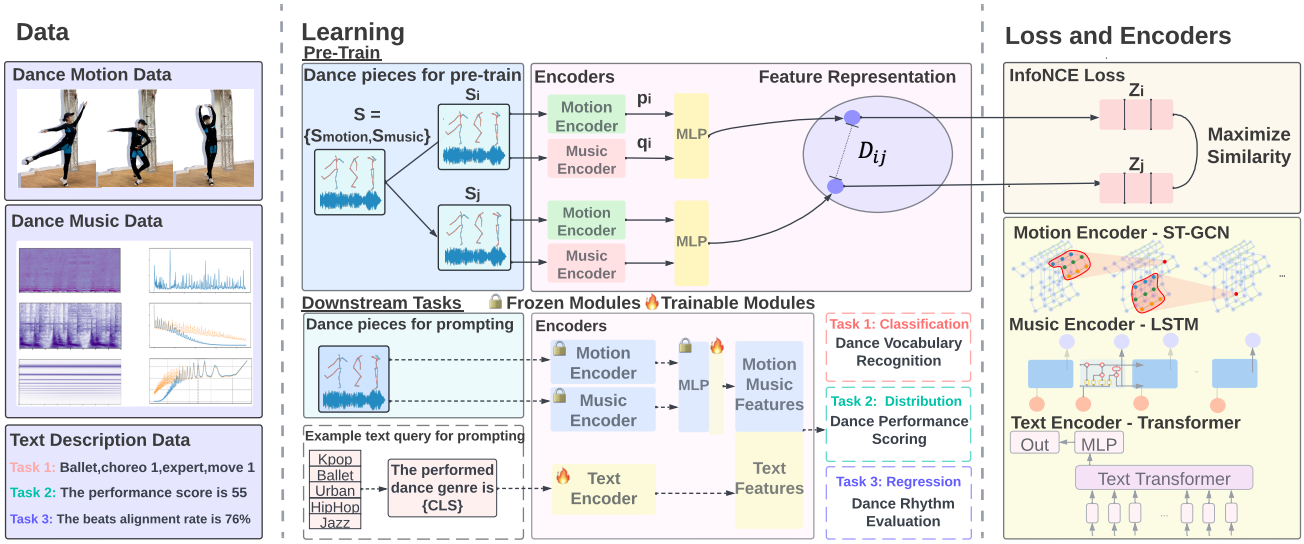

Figure 2: The proposed self-supervised framework. The augmented pair of dance pieces is mapped onto the lower-dimensional latent space by using ST-GCN and LSTM encoders and the contrastive InfoNCE loss. As for the downstream tasks, Transformer-based text prompt tuning is applied to make the pre-trained knowledge adapt to three dance assessment tasks effectively.

classification and proved that various compositions of data argumentation could result in good feature representations. Most existing self-supervised methods are designed for visual tasks, and models that use human 3D-skeleton data as input are less explored. (Thoker, Doughty, and Snoek 2021) uses human 3D-skeleton data as input and explores various skeleton augmentations to enhance the performance of human action recognition. In this paper, we deploy contrastive learning to represent human 3D-skeleton dance motions and music features and perform three downstream tasks.

Prompt Tuning As self-supervised learning strategy has gained massive success in both NLP and CV tasks, the investigation of stronger generalization ability to adapt the pre-trained model to downstream tasks even in the few-shot or zero-shot settings has become a popular research topic among the communities such as GPT-3 (Brown et al. 2020) and CLIP (Radford et al. 2021). Prompt tuning is one of the most popular methods to achieve the desired generalization ability. Inspired by recent success on prompting with vision and language models (Jia et al. 2022; Lester, Al-Rfou, and Constant 2021; Radford et al. 2021), we extend the prompt tuning technologies to our dance-music multimodal model to address the challenges of adapting our pre-trained dance model to diverse unseen dance motions and music.

DanceMVP

Problem Formulation

As shown in Fig. 2, we propose a self-supervised learning framework to perform multi-task primitive-based dance performance assessment. Our goal is to evaluate the performance of the input dance sequences. The evaluation tasks are: (i) dance vocabulary recognition: recognition of dance motion and music primitives, genres, choreographies, and expertise levels; (ii) dance performance scoring: prediction of the dance score distribution and (iii) dance rhythm evaluation: regression of the music and motion alignment rate.

The whole process could be divided into two stages: we first deploy ST-GCN and LSTM encoder to contrastively learn the dance motion and music feature representation with InfoNCE loss using unlabelled data. The model input is a set of sequences $\mathcal{S} = \{\mathcal{S}_{motion}, \mathcal{S}_{music}\}$ of dance motion and music. We then carry out the multi-task downstream knowledge transfer on unseen data using the transformer-based text prompt tuning methodology.

Pre-training

Eight-Beats Primitive Segmentation Before training, we first use an Eight-Beats segmentation method to recognize the motion and music primitives of the input dance sequences. Eight-beats is an approach of breaking down and counting dance moves (Royal Academy of Dance 2020). We follow this rule and divide the dance sequences into motion primitives based on every eight beats of the corresponding music. Fig. 3 shows the representation of our Eight-Beats segmentation strategy. As we could observe from Fig. 3(b), a sequence of music samples is divided into three groups of eight beats (pink dash line). The beats are detected by the audio signal processing library - *librosa* (McFee et al. 2015). The dance motions are also divided into three groups of motion primitives according to the musical beats as in Fig. 3(c). According to Fig. 3, the movements shown in the first group are opposite to the third group, which indicates this Eight-Beats method is capable of segmenting meaningful dance motions based on the choreography and musical content.

Motion and Music Encoders In this paper, the dance motion features are extracted by implementing ST-GCN (Yan, Xiong, and Lin 2018), which is a Graph-Based Neural Net-

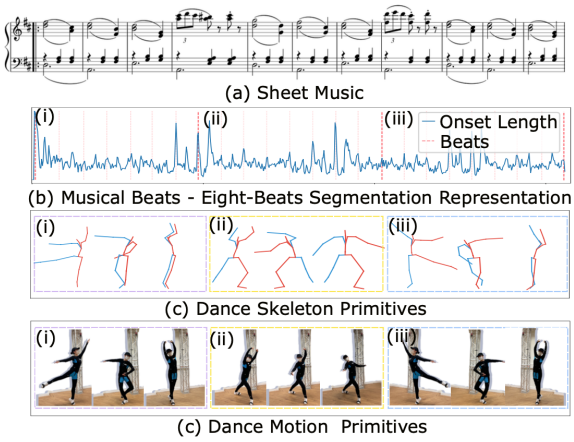

Figure 3: Eight-Beats segmentation. The dance music (b) and the motion (c) are divided into 3 groups of primitives, reflecting the choreography pattern and music rhythm (a).

work that learns the spatial and temporal features from human motion. The graph's input is the sequence of dance motion $\boldsymbol{m}_i^k$, where i is the index of the body joints and k is the number of the frame. We denote this encoder as $\phi(\cdot)$:

$$\boldsymbol{p}_i = \textit{ST-GCN}(\boldsymbol{m}_i^k) = \phi(\boldsymbol{m}_i^k)$$

For music feature learning, we first pre-process the raw music to extract the acoustic features, as the acoustic feature is expected to be more informative and non-redundant than the raw music. The acoustic features are retrieved by using *librosa* (McFee et al. 2015). Specifically, there are 5 features categories $\mathcal{F}(\boldsymbol{a}_i) = \{\boldsymbol{a}_0, \boldsymbol{a}_1, ..., \boldsymbol{a}_4\}$ extracted in this study: the Mel-frequency cepstral coefficients (MFCC), MFCC-delta, constant-Q chromagram, tempogram and onset strength. After the acoustic feature extraction, we adopt LSTM $\psi(\cdot)$ to learn acoustic features as in (Nguyen, Nguyen, and Freedman 2023; Li et al. 2018).

$$\boldsymbol{q}_i = \textit{LSTM}(\boldsymbol{a}_i) = \psi(\boldsymbol{a}_i)$$

The feature embedding $\boldsymbol{z}_i$ of the motion and music is obtained by concatenating 2 encoder outputs by an MLP.

Contrastive InfoNCE Loss After obtaining the feature embedding $\boldsymbol{z}_i$ of motion and music, the InfoNCE loss (Sohn 2016) is applied as it could efficiently learn feature representations and achieve generalizability in various domains (Chen et al. 2020; Saeed, Grangier, and Zeghidour 2021). The loss compares the distance to positive samples and the distance to negative samples for each positive pair of the network output. The training is formulated based on the augmented pairs that are obtained from each sample, resulting in $2N$ data points. According to (Chen et al. 2017), apart from the positive sample, we treat the other $2(N-1)$ samples within a minibatch as negative samples. If (i, j) is a positive pair, then the loss is defined as Equation (1):

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{r=1}^{2N} 1_{[r \neq i]} \exp(\text{sim}(z_i, z_r)/\tau)} \qquad (1)$$

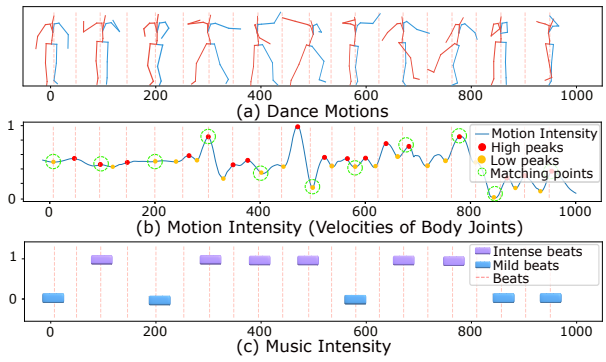

Figure 4: Beats alignment. For an expert dancer, the peaks (pink dash-line) of the motion (green circle) and music intensity mostly match each other.

where the indicator function $1_{[r \neq i]} \in \{0, 1\}$ is estimated to be 1 if $r = i$. we could conclude from Equation (1) that the loss value is low if positive samples are encoded to similar (closer) representations in latent space and negative samples are encoded to dissimilar (further) representations. The pairs are created by adpoted the data augmentation technique.

Downstream Tasks

Assessing one's dance performance requires professional multi-sensory analysis of multimodal data. We formulate this problem as three tasks: (i) dance vocabulary recognition, (ii) dance performance scoring and (iii) dance rhythm evaluation. These assessment criteria are usually adopted by various professional dance institutes, including the Royal Academic of Dance, UK (Royal Academy of Dance 2020). In the downstream phase, we present unseen testing dance pieces to the pre-trained model and evaluate each task by adopting a text prompting method.

Transformer-based Text Prompting Fig. 2 shows our proposed downstream text prompting workflow. We describe the given testing dance piece contents (e.g. types of dance genres, levels, or scores) in text, and use it as the input. We then deploy a transformer encoder to extract the feature embedding from the input raw text. The text input and the text encoder will serve as a prompter. The text features are concatenated with the motion-music features, which are then used as predictions to compare with the ground truth of each downstream task using task-agnostic loss. The pre-trained motion and music encoder parameters are frozen, and only the text encoder and the last layer of the MLP are trained during the downstream phase. The Data Section (purple) of Fig. 2 shows examples of the input text for the three evaluation tasks. For example, an input of the first task's text encoder is *Ballet, choreography 1, beginner and move 2*, which verbally describes the dance genre, choreography, expertise level and primitive class of the testing piece.

Multi-Task Dance Performance Assessment
Dance Vocabulary Recognition The dance vocabulary recognition includes the identification of the dance genres,

choreographies, expertise levels and primitives of dance sequences. We formulate this as a multi-label classification problem of different genres, choreographies, expertise levels and motion. We have 20 choreographies, 3 expertise levels, and 3 primitive pieces, resulting in $20 \times 3 \times 3 = 180$ classes. We then implement the cross-entropy loss to determine the dance vocabulary during the network learning.

Dance Performance Scoring We formulate our second task as a score distribution prediction problem. Each dance sequence in our dataset is also labeled with a performance score by professional dancers as ground truth. Instead of predicting scores, our approach learns the mapping between the features and a score distribution that involves the aleatoric uncertainty, aiming to reduce any subjective impact during the labeling process (Zhang et al. 2021). Gaussian distribution is used to model the estimated score and predict the mean and variance of the distribution. The distribution optimization could be considered as finding the maximum log-likelihood value of the target distribution, which is equal to minimizing its negative value. Hence the loss becomes:

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{\alpha}{\sigma(x_i)^2}||y_i - \mu(x_i)||^2 + \beta \log \sigma(x_i)^2\right\}$$

where x and y are the input data and the score distribution, μ and σ are the mean and variance of the distribution, and α and β are the weights of the uncertainty and constant part.

Dance Rhythm Evaluation Assessing dance performances should not ignore the music, as the motion should match the musical rhythm. In this study, we use the beats alignment rate (BAR) to represent this correlation between the music and motions. We recruit professional dancers to label the music intensity based on the choreography (Fig. 4(c)) and calculate the motion intensity based on the sum of the human body joints velocities (green circles in Fig. 4(b)). From Fig. 4 we can observe that the music and motion intensity of an expert dancer approximately match each other, further proving their strong correlation. The motion and music intensity are matched if the difference between each motion and music intensity peak is not greater than 0.2 seconds. We calculate the percentage of the matching intensity peaks for each dance sequence as ground truth. Then a Mean Squared Error loss is used for network learning.

ImperialDance Dataset

We developed ImperialDance, which comprises 69,300 seconds of dance motions and music, 5 genres, 20 pieces of music and 20 choreographies with dancers of 3 different expertise levels. Table 1 shows the comparison between our ImperialDance with the other latest public datasets. Two advantages of our dataset could be summarized as (i) Our ImperialDance is the first dataset that describes the expertise level differences of dancing, which promotes the research on dancing skill assessment. (ii) Compared to other dance-music datasets that only provide one sample of the dance motion per choreography, we provide 100 repeating samples for each class (per music, per genre, per choreography and expertise level). This will benefit the feature learning of particular dance motions and level progression modeling.

Dataset	Levels	Choreo	Samples	Subjects	Seconds
JIGSWAS	✓	-	30	7	515
BEST	✓	-	100	500	25000
DanceNet	✗	2	1	2	3472
Dance w/ Melody	✗	40	1	-	5640
GrooveNet	✗	4	1	1	1380
AIST++	✗	101	2	10	18694
ImperialDance	✓	20	100	5	69300

Table 1: Comparision between our ImperialDance and other state-of-the-arts, including BEST (Doughty, Mayol-Cuevas, and Damen 2019), JIGSWAS (Gao et al. 2014), DanceNet (Zhuang et al. 2022), Dance with Melody (Tang, Jia, and Mao 2018), GrooveNet (Alemi, Françoise, and Pasquier 2017) and AIST++ (Li et al. 2021).

Experiments

We conduct three experiments to demonstrate: (i) Features Learning Evaluation: the validity of our proposed strategy on dance motion-music feature representation for different dance genres, choreographies, expertise levels and motion primitives. (ii) Downstream Tasks via Prompting: the effectiveness of our text prompting method on the proposed three assessment tasks. (iii) Baselines and Ablation Analysis: the results of our method in comparison with different studies.

Implementation Details

We use 10 different dance choreographies data that cover 5 different dance genres for pre-training. Each choreography contains 100 repeating samples per expertise level and per motion primitive. All the sequences are processed into identical lengths of 10 seconds, which are then segmented into three pieces (3 seconds each) according to the Eight-Beats segmentation method, resulting 90,000 training dance pieces ($90,000 = 10 \times 100 \times 3 \times 3$). As for the downstream tasks, we use the other 5 dance choreographies that belong to 5 different genres for testing, resulting in 4,500 pieces ($4,500 = 5 \times 100 \times 3 \times 3$). The model is trained in PyTorch using RTX 3080 GPUs, with a batch size of 16 and a learning rate of 3×10^{-4}. The pre-train and downstream training times are about 8 and 5 hours. The downstream inference time is 0.3 seconds with 32 batch size. We use (i) classification accuracy, (ii)log-likelihood value and (iii) MSE loss as the metrics to evaluate the 3 performance assessment tasks.

Dance Features Learning Evaluation

Pre-trained Feature Representation As described in the Implementation Detail, we pre-train a model using 90,000 dancing pieces from our proposed ImperialDance dataset. The obtained latent features are mapped onto 2D space using t-SNE for visualization. For clarity, we show the results in Fig. 5(a) of 3 dance genres. Each dance genre includes 1 choreography, each choreography contains 3 different expertise levels data, each sequence of particular expertise level is segmented into 3 primitive pieces, and each piece per expertise level contains 100 samples, resulting 2700 data points ($2700 = 3 \times 1 \times 3 \times 3 \times 100$). Three observations could be

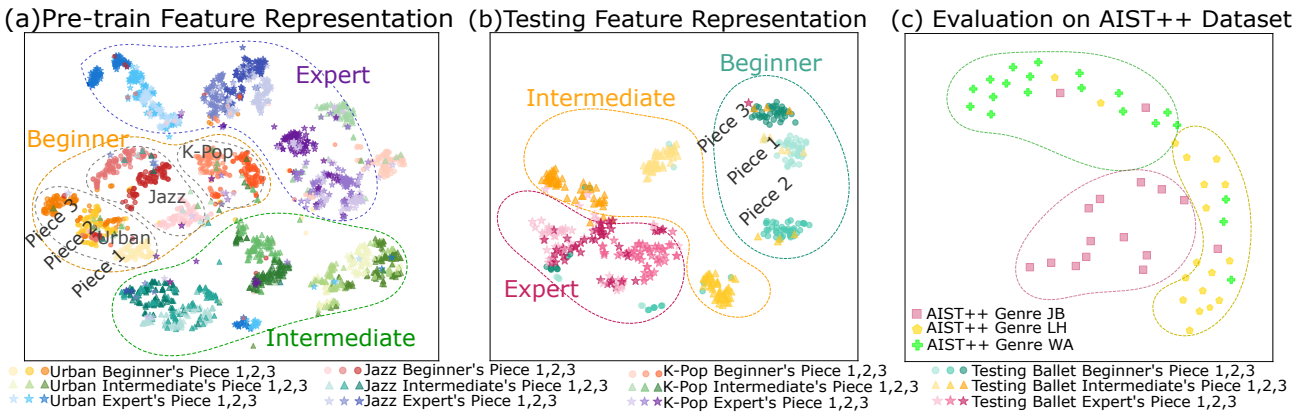

Figure 5: Feature learning analysis. (a) Pre-train features. 3 different dance genres, 3 expertise levels and 3 motion primitives could be separated simultaneously. (b) Downstream tasks features. The testing choreographies that are unseen during pre-train are classified. 3 expertise levels for 3 motion primitives are also separated. (c) AIST++ dataset features.

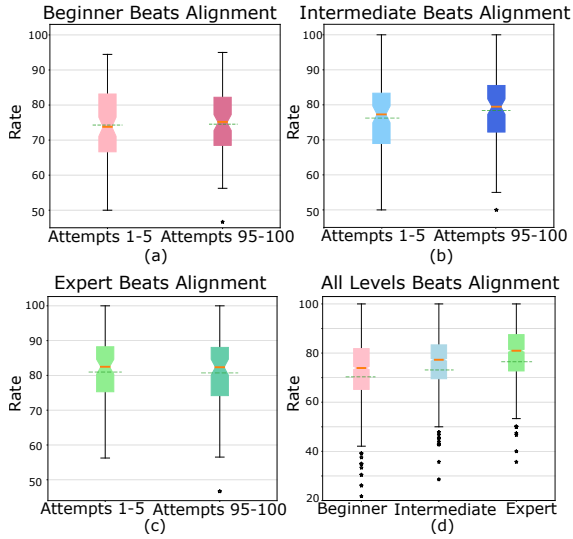

Figure 6: Progression modeling. The BAR increases for beginners and intermediates, and the experts maintain the same after 100 times of practice of dance ((a)-(c)). For all choreographies, the level increases, and the BAR increases ((d)).

summarized from Fig. 5(a): (i) The method can separate different dance genre data in the feature space. The 3 selected dance genres are surrounded by dotted lines in 3 colors (orange, purple, green). (ii) The method is also capable of separating different expertise levels of each choreography. For example, the Expert data of Urban, Jazz and K-Pop dance are classified into 9 clusters (blue stars, navy blue stars and purple stars), which are far away from other levels of data points (scatters and triangles). (iii) The method is also able to separate different motion primitive pieces for each genre, choreography and expertise level. For example, the beginner data of Urban dance is separated into 3 classes (yellow scatters that are surrounded by grey dotted lines).

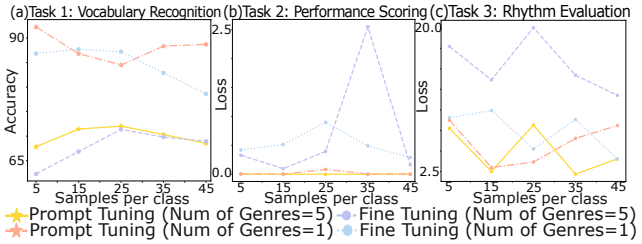

Figure 7: Comparing prompt tuning v.s. fine-tuning. Our method outperforms fine-tuning 45% with fewer training samples (5 per class) across three tasks. Also, when the number of unseen testing choreographies increases (from 1 to 5), the fine-tuning performance decreases avg. 37%, while our prompt tuning can maintain a decent performance.

Testing Feature Representation We then use the unseen dance choreographies to test our pre-trained model. The testing data includes 5 different dance genres. There is 1 choreography for each genre. For clarity, we show the results from 1 unseen dance choreography (Ballet). Fig. 5(b) shows the Ballet testing feature representation. Two observations could be made from this result: (i) Our method shows the ability to separate different expertise levels for the testing choreography. As we can see from Fig. 5(b), for this Ballet choreography, 3 expertise levels (green, yellow and pink) are separated. (ii) Our method is also capable of separating different motion primitive pieces of a dance sequence. For the Ballet beginner data (green), the 3 motion primitives (light green, green, dark green) are clustered in 3 different groups. Also, we tested the pre-trained model with the state-of-the-art dance-music AIST++ dataset aiming to show the generalizability of our approach and our dataset. As we can see from the Fig. 5(c), 3 dance genres of AIST++ that are unseen by the model are also separated in the feature space.

Progression Modeling Our ImperialDance dataset contains 100 repeating samples for each dance choreography at each expertise level, allowing us to track the dance perfor-

Method	Task 1 ↑	Task 2 ↓	Task 3 ↓
SimCLR	31.85 ± 0.59	0.069 ± 0.01	19.23 ± 0.01
MoCo	16.29 ± 2.63	2.85×10^{-5}	6.60 ± 18.43
SVT	55.08 ± 0.56	0.013	10.89 ± 2.09
CrossEntropy	12.76 ± 0.45	1.06×10^{-3}	15.65 ± 0.07
SupCon	30.72 ± 0.86	0.199	4.20 ± 0.76
CotrastiveDance	66.85 ± 0.19	0.097 ± 0.01	13.65 ± 0.77
HappyFeet	15.63 ± 0.43	$\mathbf{2.52 \times 10^{-5}}$	16.49 ± 0.41
Dance w/ Melody	21.44	2.83×10^{-4}	5.30 ± 0.03
Ours	$\mathbf{71.43 \pm 0.71}$	2.56×10^{-4}	$\mathbf{2.52 \pm 0.73}$

Table 2: Baselines. Our method outperforms state-of-the-art self-supervised and supervised methods (first 5 rows) by at least 44% and 38%, and outperforms the latest dance-music studies by at least 7% and 52% for Task 1 and Task 3.

Motion-Music-Loss	Task 1 ↑	Task 2 ↓	Task 3 ↓
Tx-Tx-InfoNCE	54.46	0.023	14.48 ± 20.14
Tx-LSTM-InfoNCE	29.94 ± 0.47	0.014	16.89 ± 5.07
Res18-Res18-InfoNCE	43.49 ± 0.44	2.86×10^{-3}	4.06 ± 4.60
Res50-Res50-InfoNCE	39.84 ± 2.72	5.79×10^{-3}	5.33 ± 1.53
Res18-LSTM-InfoNCE	55.21 ± 0.21	1.37×10^{-3}	$\mathbf{2.08 \pm 0.51}$
Res50-LSTM-InfoNCE	60.68 ± 0.58	1.42×10^{-3}	2.44 ± 0.12
ST-GCN-Res18-InfoNCE	69.01 ± 1.16	2.72×10^{-4}	2.79 ± 1.10
ST-GCN-Res50-InfoNCE	64.32 ± 0.79	0.031	3.44 ± 1.15
ST-GCN-LSTM-SupCon	67.61 ± 1.09	0.0060	5.51 ± 0.05
Ours	$\mathbf{71.43 \pm 0.71}$	$\mathbf{2.56 \times 10^{-4}}$	2.52 ± 0.73

Table 3: Ablation Study 1. The proposed ST-GCN-LSTM combination with InfoNCE loss outperforms others.

mance quality over these 100 trials. We use the musical beats alignment rate (BAR) as the evaluation criterion to identify whether there is an improvement after 100 times of practice. We respectively obtain the average beats alignment rate of the first and last 5 dance sequences for the Beginner, Intermediate and Expert levels (Fig. 6(a)-(c)) for all 20 choreographies. The gap between the first and last 5 sequences for each level is considered as the progression. Similarly, the average beats alignment rate for the Beginner, Intermediate and Expert across all 100 sequences for each choreography are also obtained (Fig. 6(d)). Four observations could be summarized from Fig. 6: (i) For Beginner and Intermediate, the mean and median BAR of the last 5 sequences are higher than the first 5 (Fig. 6(a) and (b)), which indicates our method is capable of modeling people's dance capability improvements after practising. (ii) By comparing Fig. 6(a) and (b), we could identify that the progression made by Intermediate dancers is larger than the Beginner. This is expected as the Intermediate dancers learn to dance to the musical beats faster than the Beginners. Thus after 100 practising, the Intermediate made more improvement than the Beginner. (iii) There is no progression identified for the Experts, as shown in Fig. 6(c). It is obvious because the expert dancers have already possessed the ability to ensure a high BAR all the time. (iv) As the level increases (Beginner to Expert), the mean and median value of BAR also increases (Fig. 6(d)).

Method	Task 1 ↑	Task 2 ↓	Task 3 ↓
Prompting (g=5)	$\mathbf{71.43 \pm 0.71}$	$\mathbf{2.56 \times 10^{-4}}$	$\mathbf{2.52 \pm 0.73}$
Fine-tuning (g=5)	66.85 ± 0.19	0.097	13.65 ± 0.77
Prompting (g=1)	86.83	$\mathbf{2.91 \times 10^{-4}}$	$\mathbf{3.00 \pm 0.03}$
Fine-tuning (g=1)	$\mathbf{87.69 \pm 0.40}$	0.51 ± 0.07	9.93 ± 0.01

Table 4: Ablation Study 2. Our method outperforms the fine-tuning across all three tasks when there are 5 different unseen dance genres (abbreviated as g) presented to the model.

Baselines and Ablation Studies

Table 2 shows the comparison of our proposed self-supervised method against (i) latest self-supervised (Chen et al. 2020; He et al. 2020; Ranasinghe et al. 2022) and supervised learning (He et al. 2016; Khosla et al. 2020) approaches (first 5 rows), (ii) latest dance performance assessment studies (5th-6th rows), and (iii) dance-music synthesis method (last row). Our method outperforms others for Task 1 and 3 and obtains the third best for Task 2. Standard deviation values that are less than 0.01 in Table 2, 3 and 4 are omitted for clarity.

We also conduct 2 ablation studies which include (i) Pre-trained Alternatives: using different encoders and contrastive losses combination and (ii) Downstream Prompt Tuning v.s. Fine-Tuning: comparing the downstream tasks performance between the proposed text prompting method and traditional logistic regression fine-tuning. Table 3 shows the first ablation study result. As we can summarize from Table 3, the combination of ST-GCN and LSTM encoders together with InfoNCE loss shows the best performance in representing the motion and music features against other alternatives (Tx stands for transformer). Secondly, Table 4 shows our transformer-based text prompter achieves the best performance for most of the tasks, especially when there are 5 different unseen dance choreographies presented to the model. We can also observe from Fig. 7 that when using fewer training samples, our prompting method outperforms the fine-tuning in all three tasks. One limitation is that the prompt tuning takes longer than the fine-tuning, as the Transformer has a large network size. A failure case would happen if the dance choreographies are similar.

Conclusion

We formulate a novel multi-task primitive-based dance motion analysis problem and present a self-supervised learning framework to solve it. We also propose a text transformer prompting method to interpret the unseen dance sequences, performing 3 dance performance assessment tasks with decent results. We also constructed the ImperialDance dataset that includes diverse expertise levels of dance motions and a significant amount of samples per class. The method is capable of not only evaluating human dance levels based on different dance music, genres, choreographies and motion primitives qualitatively and quantitatively but also monitoring the dancers' level improvements. Our method coule be extended to various applications such as motion generation.

Ethics Statement

Our experimental protocol has been granted Ethics Approval from Imperial College's ethics board. All collected data in the ImperialDance have associated informed signed consent forms. The dataset will be made publicly available in an anonymised form.

References

Alemi, O.; Françoise, J.; and Pasquier, P. 2017. GrooveNet: Real-time music-driven dance movement generation using artificial neural networks. *networks*, 8(17): 26.

Alexiadis, D. S.; and Daras, P. 2014. Quaternionic Signal Processing Techniques for Automatic Evaluation of Dance Performances From MoCap Data. *IEEE Transactions on Multimedia*, 16(5): 1391–1406.

Bertasius, G.; Soo Park, H.; Yu, S. X.; and Shi, J. 2017. Am I a baller? Basketball performance assessment from first-person videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2177–2185.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chan, J. C.; Leung, H.; Tang, J. K.; and Komura, T. 2011a. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Transactions on Learning Technologies*, 4(2): 187–195.

Chan, J. C.; Leung, H.; Tang, J. K.; and Komura, T. 2011b. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Transactions on Learning Technologies*, 4(2): 187–195.

Chen, K.; Tan, Z.; Lei, J.; Zhang, S.-H.; Guo, Y.-C.; Zhang, W.; and Hu, S.-M. 2021. ChoreoMaster: Choreography-Oriented Music-Driven Dance Synthesis. *ACM Trans. Graph.*, 40(4).

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Sun, Y.; Shi, Y.; and Hong, L. 2017. On sampling strategies for neural network-based collaborative filtering. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 767–776.

Doughty, H.; Mayol-Cuevas, W.; and Damen, D. 2019. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7862–7871.

Gao, Y.; Vedula, S. S.; Reiley, C. E.; Ahmidi, N.; Varadarajan, B.; Lin, H. C.; Tao, L.; Zappella, L.; Béjar, B.; Yuh, D. D.; et al. 2014. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision (ECCV)*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 18661–18673. Curran Associates, Inc.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, B.; Zhao, Y.; Zhelun, S.; and Sheng, L. 2022. Danceformer: Music conditioned 3d dance generation with parametric motion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 1272–1279.

Li, R.; Yang, S.; Ross, D. A.; and Kanazawa, A. 2021. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13401–13412.

Li, Z.; Wang, H.; Zhao, M.; Li, W.; and Guo, M. 2018. Deep Representation-Decoupling Neural Networks for Monaural Music Mixture Separation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Liu, D.; Li, Q.; Jiang, T.; Wang, Y.; Miao, R.; Shan, F.; and Li, Z. 2021. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9522–9531.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 18–25. Citeseer.

Nguyen, V. D.; Nguyen, Q. H.; and Freedman, R. G. 2023. Predicting Perceived Music Emotions with Respect to Instrument Combinations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 16078–16086.

Parmar, P.; and Morris, B. T. 2019. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 304–313.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ranasinghe, K.; Naseer, M.; Khan, S.; Khan, F. S.; and Ryoo, M. 2022. Self-supervised Video Transformer. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition*.

Ren, X.; Li, H.; Huang, Z.; and Chen, Q. 2020a. Self-supervised dance video synthesis conditioned on music. In *Proceedings of the 28th ACM International Conference on Multimedia*, 46–54.

Ren, X.; Li, H.; Huang, Z.; and Chen, Q. 2020b. Self-Supervised Dance Video Synthesis Conditioned on Music. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 46–54. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Royal Academy of Dance, A., Examinations. 2020. QCF Examinations Information, Rules and regulations - Royal Academy of Dance.

Saeed, A.; Grangier, D.; and Zeghidour, N. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3875–3879. IEEE.

Sharma, Y.; Bettadapura, V.; Plötz, T.; Hammerla, N.; Mellor, S.; McNaney, R.; Olivier, P.; Deshmukh, S.; McCaskie, A.; and Essa, I. 2014. Video based assessment of OSATS using sequential motion textures. Georgia Institute of Technology.

Siyao, L.; Yu, W.; Gu, T.; Lin, C.; Wang, Q.; Qian, C.; Loy, C. C.; and Liu, Z. 2022. Bailando: 3D Dance Generation by Actor-Critic GPT With Choreographic Memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11050–11059.

Sjödahl Hammarlund, C.; Jarnlo, G.-B.; Söderberg, B.; and Persson, B. 2002. Kinematic and kinetic gait analysis in the sagittal plane of trans-femoral amputees before and after special gait re-education. *Prosthetics and orthotics international*, 26: 101–12.

Sohn, K. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29.

Sumer, O.; Dencker, T.; and Ommer, B. 2017. Self-Supervised Learning of Pose Embeddings From Spatiotemporal Relations in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Tang, T.; Jia, J.; and Mao, H. 2018. Dance with Melody: An LSTM-autoencoder Approach to Music-oriented Dance Synthesis. *Proceedings of the 26th ACM international conference on Multimedia*.

Tang, Y.; Ni, Z.; Zhou, J.; Zhang, D.; Lu, J.; Wu, Y.; and Zhou, J. 2020. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9839–9848.

Thoker, F. M.; Doughty, H.; and Snoek, C. G. 2021. Skeleton-contrastive 3D action representation learning. In *Proceedings of the 29th ACM International Conference on Multimedia*, 1655–1663.

Xu, J.; Rao, Y.; Yu, X.; Chen, G.; Zhou, J.; and Lu, J. 2022. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2949–2958.

Yan, S.; Xiong, Y.; and Lin, D. 2018. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

Zhang, B.; Chen, J.; Xu, Y.; Zhang, H.; Yang, X.; and Geng, X. 2021. Auto-Encoding Score Distribution Regression for Action Quality Assessment. *arXiv preprint arXiv:2111.11029*.

Zhang, F.; and Demiris, Y. 2022. Learning garment manipulation policies toward robot-assisted dressing. *Science robotics*, 7(65): eabm6010.

Zhong, Y.; Zhang, F.; and Demiris, Y. 2023. Contrastive Self-Supervised Learning for Automated Multi-Modal Dance Performance Assessment. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhuang, W.; Wang, C.; Chai, J.; Wang, Y.; Shao, M.; and Xia, S. 2022. Music2dance: Dancenet for music-driven dance generation. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2): 1–21.