M2Beats: When Motion Meets Beats in Short-form Videos

Dongxiang Jiang[†] , **Yongchang Zhang**[†] , **Shuai He** , **Anlong Ming**[*]
School of Computer Science (National Pilot Software Engineering School),
Beijing University of Posts and Telecommunications
{jiangdx, zhangyongchang, hs19951021, mal}@bupt.edu.cn

Abstract

In recent years, short-form videos have gained popularity and the editing of these videos, particularly when motion is synchronized with music, is highly favored due to its beat-matching effect. However, detecting motion rhythm poses a significant challenge as it is influenced by multiple factors that make it difficult to define using explicit rules. While traditional methods attempt to define motion rhythm, they often yield unsatisfactory results. On the other hand, learning-based methods can extract motion rhythm without relying on explicit rules but require high-quality datasets. Unfortunately, existing datasets simply substitute music rhythm for motion rhythm which are not equivalent. To address these challenges, we present the motion rhythm dataset *AIST-M2B*, which is annotated with meticulously curated motion rhythm labels derived from the profound correlation between motion and music in professional dance. We propose a novel network architecture called *M2BNet* that is specifically trained on *AIST-M2B* to effectively extract intricate motion rhythms by incorporating both human body structure and temporal information. Additionally, we introduce a pioneering algorithm for enhancing motion rhythm synchronization with beats. Experimental results substantiate the superior performance of our method compared to other existing algorithms in the domain of motion rhythm analysis. Our code is available at https://github.com/mRobotit/M2Beats.

1 Introduction

In recent years, there has been a rapid growth of short-form videos. An increasing number of users share dance videos paired with music, creating a pleasurable rhythmic experience. However, raw videos often lack coherence and professional editing is both time-consuming and costly. The demand for automated enhancement of video rhythm has become imperative, while research on enhancing video motion rhythm remains in its nascent stages.

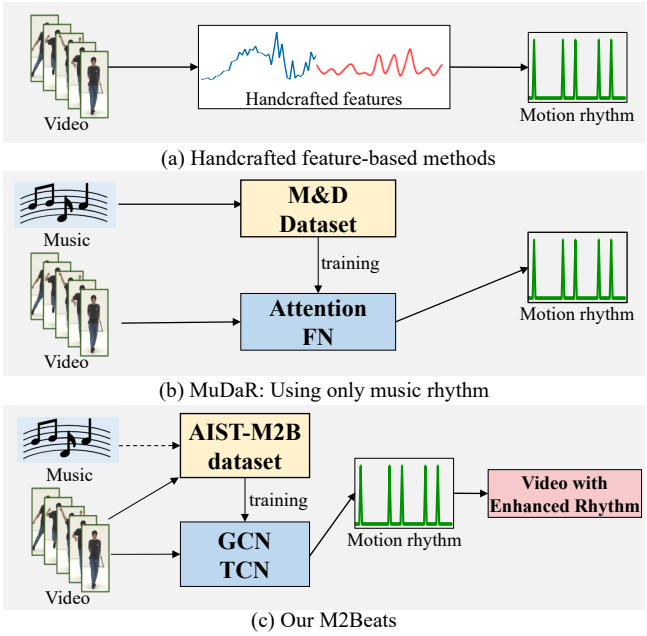

Figure 1: Pipelines of motion rhythm extraction methodologies.

Humans have consistently sought beauty [He *et al.*, 2022] and harmony [Tchemeube *et al.*, 2023], several research studies aim to define motion rhythm. Dance analysis studies, such as [Kim *et al.*, 2003; Shiratori *et al.*, 2006a; Chu and Tsai, 2011], concur that key dance poses occur during stops or turns in the performer's movement trajectories. Recent empirical research, exemplified by [Chen *et al.*, 2021; Bellini *et al.*, 2018], employs handcrafted features to extract motion rhythm, such as identifying the troughs of velocity curves for body movements, as shown in figure 1 (a). However, existing methods face challenges due to disagreements regarding the definition of motion rhythm, which is inherently difficult to detect using explicit rules given its dependence on multiple factors. Consequently, these approaches are prone to yielding suboptimal performance.

Learning-based methods have emerged as a prevailing trend, showcasing remarkable generalization capabilities. The efficacy of these approaches heavily relies on the quality of the dataset employed. However, challenges arise due to

[*]Corresponding author.

the scarcity of high-quality datasets capturing motion rhythm, which necessitate costly manual annotations or encounter difficulties in automatic annotation using explicit rules. Some methods circumvent the need for annotation, such as exemplified by [Zhou *et al.*, 2023], where noise is added to professional motion and an algorithm is trained to refine amateur motion. Nevertheless, this approach compromises motion information and overlooks dance rhythm considerations, resulting in suboptimal visual perception. Furthermore, employing an end-to-end approach poses challenges for subsequent tasks. MuDaR [Yu *et al.*, 2023] addresses this issue by crawling dance videos from the web and utilizing music rhythm labels for annotating motion rhythm; however, it should be noted that motion rhythm is inherently determined by individual dancers and does not possess a one-to-one correspondence with music rhythm. The regularity of music rhythm may potentially mislead deep learning models during training.

The comparison of existing methods for motion rhythm extraction is illustrated in Figure 1. The Handcrafted feature-based methods (figure 1 (a)) involve the design of manually crafted features for extracting motion rhythm. Subsequently, MuDaR, the learning-based method based on the M&D dataset [Yu *et al.*, 2023], considers music rhythm as an indicator of motion rhythm (figure 1 (b)). To address labeling ambiguities, our proposed method automatically annotates motion rhythm by integrating both motion and music information. We present the AIST-M2B dataset for motion rhythm analysis, leveraging robust rhythmic motions to mitigate labeling uncertainties and employing a sophisticated algorithm for automatic annotation of motion rhythm (figure 1 (c)). Our approach initially generates estimations of motion rhythm based on human movement characteristics, which are further refined using music rhythms. Benefiting from the inherent strong rhythmic patterns observed in professional dance videos, our annotated motion rhythms exhibit exceptional quality. The extracted motion rhythms can be utilized to enhance video content.

In this paper, we propose a learning-based method called M2BNet to extract motion rhythm, which is trained using the AIST-M2B dataset. Previous methods commonly rely on handcrafted features such as troughs in velocity curves, which limits their generalization ability. Although the existing MuDaR [Yu *et al.*, 2023] approach employs a learning-based method for extracting motion rhythm, it overlooks the structural characteristics of the human skeleton. To address this limitation, we introduce motion feature extraction (MFE) block consisted with skeleton graph convolution network (SGCN) and temporal convolutional network (TCN) for analyzing the human skeleton and incorporating temporal information. Our contributions are summarized as follows:

- The meticulously curated annotations of our motion rhythm dataset, AIST-M2B, establish it as an indispensable resource for learning-based methodologies and facilitate comprehensive comparisons among related approaches, thereby rendering it a valuable asset in the realm of academic research.
- We propose M2BNet, the novel approach for extracting motion rhythm from videos by integrating MFE block, which demonstrates superior performance compared to previous methodologies.
- The present study proposes a novel method for enhancing the rhythm of videos. Our experimental findings in the domain of short-form video processing underscore the substantial potential inherent in this approach.

2 Related Works

2.1 Handcrafted Feature-based Methods

A precise definition of motion rhythm is essential for a mathematical approach. Some researchers (e.g., [Shiratori *et al.*, 2004; Chu and Tsai, 2011; Shiratori and Ikeuchi, 2008]) suggest that the movement trajectories of hands, feet, and center of body can represent motion rhythm. Others ([Bellini *et al.*, 2018; Davis and Agrawala, 2018]) extract local minimum in velocity curve of optical flow as motion rhythm. Additionally, [Chen *et al.*, 2021] detects foot contact time and introduces the motion kinematic curve ([Shiratori *et al.*, 2006b]) to detect motion rhythm. The advantage of these methods is simplicity of the process and fast speed.

However, motion rhythm which derives from human sensation produce a challenge in defining it with explicit rules. Due to the influence of multiple factors, existing methods have varying detection rules for motion rhythm. Mathematical approaches attempt to define motion rhythm from different perspectives but struggle to achieve sufficient consensus, exhibit unsatisfactory performance.

2.2 Learning-based Methods

Learning-based methods have demonstrated strong generalization ability and are currently a prevailing trend in the field. The accuracy and generalization of these methods heavily rely on the quality of the dataset used for training. However, manual annotation is both costly and prone to label conflicts due to individual differences in perception [He *et al.*, 2022]. To address this issue, [Yu *et al.*, 2023] leverages dance videos obtained from web crawling and utilizes music rhythm labels as annotations for motion rhythm during model training, their M&D dataset effectively mitigates the high cost associated with manual annotation.

However, the M&D dataset has certain limitations. At the dataset level, music rhythm cannot replace motion rhythm as they do not have a one-to-one correspondence. The regularity of music rhythm may mislead deep learning models during training. These challenges become more pronounced when considering amateur dance videos on social media platforms. At the network level, while MuDaR [Yu *et al.*, 2023] employs a attention-based fully connected layer (FC) approach to extract motion rhythm, it overlooks the structure of human skeleton. In contrast, our approach incorporates a more coherent method for defining motion rhythm labels by considering both spatial characteristics and temporal information of the human skeleton to enhance performance.

2.3 Audio-Video Alignment

The term "Audio-Video Alignment" pertains to the adjustment of audio and video timing in relation to each other.

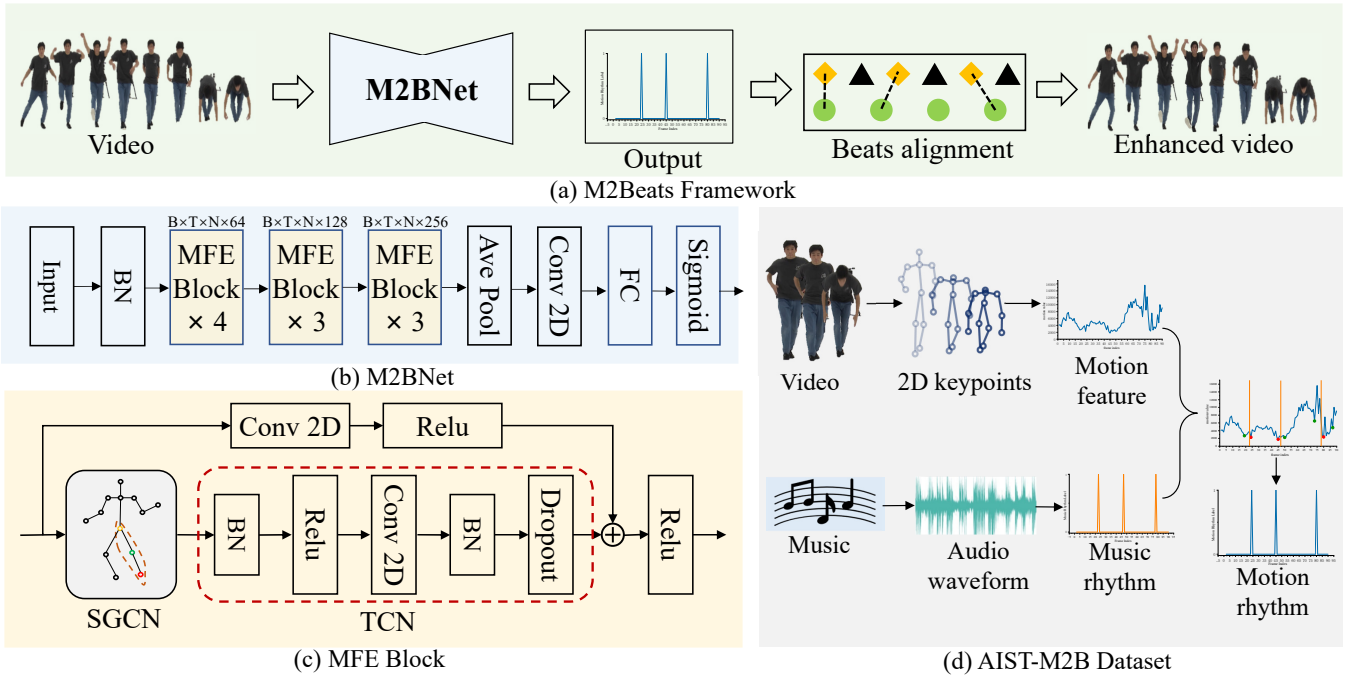

Figure 2: Pipeline of M2Beats and dataset construction process. (a) Given a sequence of motions, M2BNet extracts its rhythm and aligns it with the music rhythm, enhancing the video's overall rhythm. This alignment is visualized using different colors: yellow for motion rhythm, green for music rhythm, and black for non-rhythmic elements. (b) M2BNet consists of multiple MFE modules. (c) MFE Block mainly includes a SGCN, a TCN, and a residual structure. (d) The coarse motion rhythm is initially extracted based on movement characteristics, and then refined using music rhythm.

Early works like [Bredin and Chollet, 2007; Sargin *et al.*, 2007] used canonical correlation analysis (CCA) for synchronization prediction. Some methods (e.g., [Owens and Efros, 2018; Chung and Zisserman, 2017; Halperin *et al.*, 2019]) train deep learning models for video alignment. [Wang *et al.*, 2020] introduces random misalignment into a dance dataset and extracts motion and video features for modeling purposes in order to align videos. Nevertheless, the generalization of their end-to-end algorithm is constrained by the dataset used, making it challenging for other processes to utilize the extracted information.

In this paper, we propose a method for enhancing the rhythm of short-form videos by leveraging the extracted motion rhythm. Our approach involves dynamically synchronizing video frames with the music rhythm based on the underlying motion patterns, thereby establishing a seamless integration between motion and music.

2.4 Evaluation Metrics

In the field of motion beats, there is a lack of established mainstream evaluation metrics. However, we can draw inspiration from music evaluation metrics (CMLc, CMLt, AMLc, AMLt) used in madmom [Böck *et al.*, 2016] for extracting music rhythm. It should be noted that the evaluation routines in madmom accommodate metrical ambiguities such as half-beat time or triple and 1/3rd variations. Given the relatively random distribution of motion rhythm, excessively stringent evaluation metrics may not be suitable. Therefore, we propose adopting precision and recall as the evaluation metrics for model predictions on the motion rhythm dataset. F-measure and Cemgil are used to assess the similarity between the predicted beats and the dataset's beats. We use them as metrics for evaluating motion annotations.

3 M2Beats Framework

We present AIST-M2B capitalizing on the strong correlation between music and motion. Then we train M2BNet using the AIST-M2B dataset, consisting with MFE blocks for extracting temporal and spatial information. We propose an automated method that utilizes assessed motion rhythm to generate rhythm-enhanced videos. The overall framework denoted as figure 2 (a).

3.1 AIST-M2B Dataset

Building upon the AIST++ dataset, which comprises hundreds of dance motions and music compositions, we present AIST-M2B. While AIST++ focuses on generating dance motions that are well-aligned with input music, our AIST-M2B exploits their inherent correlation to calibrate potential motion rhythm labels. As such, it can serve as a foundational resource for learning-based algorithms aimed at refining or detecting motion rhythms.

We have employed a novel approach to obtain motion rhythm labels of high quality by exploiting the inherent correlation between dance movements and accompanying music in the videos. In contrast to [Yu *et al.*, 2023], which directly

utilizes music rhythm as a proxy for motion rhythm, our proposed method offers more dependable annotations for capturing accurate motion rhythms.

The dataset annotation process is illustrated in Figure 2 (d). We employ a filtering technique to extract the motion rhythm, capitalizing on the strong correlation between music and motion. By extracting 2D keypoints that represent motion information, we obtain candidate motion rhythm and identify onsets as music rhythm. Finally, we refine the candidate motion rhythm by incorporating onsets to derive accurate motion rhythm labels for the dataset.

Data Preprocessing. To explore the potential rhythm of motion, it is crucial to identify the pertinent information that can effectively represent dance motion. 2D keypoints enable identification of body parts and facilitate capturing subtle movements. To represent motion information in accordance with COCO format standards, we extract 17 keypoints on an individual using YOLOX [Ge *et al.*, 2021] for detecting bounding boxes around people and employ ResNet50 [He *et al.*, 2016] for keypoint detection.

The representation of music rhythm is achieved through the utilization of onsets, which serve as indicators for sudden increases in music volumes. In accordance with [Böck and Widmer, 2013], spectrograms are initially obtained by applying time-windowed FFT to the audio signals.We utilized the method provided by LibROSA [McFee *et al.*, 2015] extracting onsets with default parameters.

Motion Rhythm Annotation. Manual labeling of motion rhythm is susceptible to subjective variations, making it challenging to ensure annotators possess expertise in movement. To address this issue, we propose an automated approach for obtaining accurate labels of motion rhythm. The process of creating these labels involves two steps: extracting initial approximations of the motion rhythm and subsequently refining them using the corresponding musical rhythms.

1) Coarse Motion Rhythm. In general, motion that undergoes sudden deceleration during rapid movements is more likely to be selected as a motion rhythm. Studies focusing on dance analysis, such as [Kim *et al.*, 2003; Shiratori *et al.*, 2006a; Chu and Tsai, 2011], concur that significant dance poses occur at points of cessation or changes in the performer's movement trajectories. The motion data in AIST++ is sourced from professional dancers and exhibits a pronounced sense of rhythm. This observation motivates us to directly extract preliminary motion rhythms from AIST++.

Firstly, the velocity of human motion V is computed based on the 2D keypoints.

$$V_{i,j} = \frac{\sqrt{(K^x_{i+1,j} - K^x_{i,j})^2 + (K^y_{i+1,j} - K^y_{i,j})^2}}{\Delta t}, \quad (1)$$

where $(K^x_{i,j}, K^y_{i,j})$ denotes the locations of the j-th keypoint in the i-th frame, $V_{i,j}$ denotes the velocity of the j-th keypoint in the i-th frame.

Then, we calculate the average velocity of all keypoints of the body for each frame, denoted as V_i. Finally, we utilize a strategy to extract coarse motion rhythm by extracting valley values. There are three restrictions for extracting the local minimum: *1*) V_i is the minimum value of speed. *2*) V_i is below the average of its surrounding values by a threshold. *3*) the interval between V_i and V_{i-1} is bigger than w. R^{mot} is the set of all V_i that satisfy the following conditions:

$$R^{mot} = \{V_i \mid C_1 \wedge C_2 \wedge C_3\}, \quad (2)$$

where the conditions C_1, C_2, and C_3 are defined as follows:

$$C_1 = \{V_i \mid \min(V_{[i-p_m, i+q_m]})\}, \quad (3a)$$

$$C_2 = \{V_i \mid V_i \leq \text{mean}(V_{[i-p_a, i+q_a]}) - \delta\}, \quad (3b)$$

$$C_3 = \{(i - i_{\text{prev}}) > w\}, \quad (3c)$$

where p_m, q_m, p_a, q_a, and w are parameters that ensure the computation executes within reasonable time intervals, condition C_1 ensures that V_i is the minimum value within a specified window. Condition C_2 asserts that V_i is below the average of its surrounding values by a threshold δ, and condition C_3 requires that index i is sufficiently spaced from the previous valley index i_{prev} by more than wait samples. The sequence defined in Eq. (2), denoted as R^{mot} contains all values of V_i satisfying these criteria. The settings of each parameter can be found in the supplementary materials.

2) Accurate Motion Rhythm. Motion rhythm extracts solely based on motion features is coarse because some of these rhythmic labels may be too subtle to be considered as motion rhythm. We utilize music rhythm to assist in refining the motion rhythm, ensuring that the corrected motion rhythm label aligns with the characteristics of motion and resonates with human perception of rhythm. To calibrate the coarse motion rhythm labels, we establish two rules that fully exploit the correlation between motion and music:

- The motion rhythm should be synchronized with the music rhythm within a specified temporal window, ensuring coherence between the two rhythms and aligning with the distinctive features of AIST++.
- If there exist multiple motion rhythm candidates within a specific range in close proximity to the music rhythm, the one that exhibits the closest resemblance to the music rhythm is selected as the motion rhythm label. This approach ensures the sparsity of motion rhythm labels, thereby guaranteeing distinctiveness among the chosen motion rhythm labels.

To provide a more comprehensive explanation of our annotation strategy, we select a video sequence in Figure 3 to visually demonstrate the annotation process. Subsequently, we have computed the average velocity of all 2D keypoints for each frame, as depicted by the blue line. The music rhythm has been obtained using the extraction method is represented by vertical orange lines. The motion rhythm candidates are represented by both green and red dots. Finally, we have calibrated these motion rhythm candidates with the music rhythm to identify the final selected ground truth points denoted by red dots. The corresponding video frames also indicate that the motion represented by the red dot is more suitable as motion rhythm than that represented by the green dot.

3.2 M2BNet

The proposed motion rhythm detection network, named M2BNet. Figure 2 (b) and (c) illustrate the overall structure

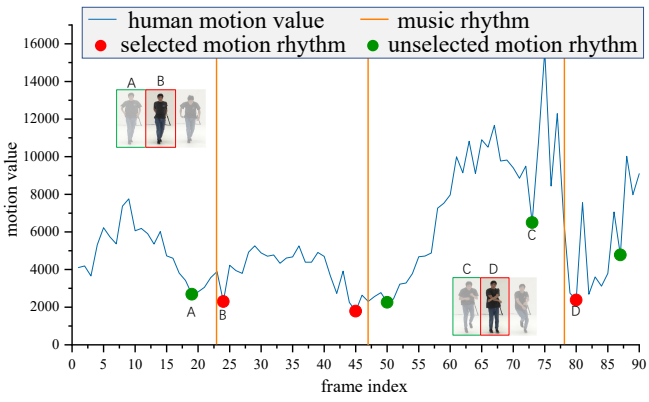

Figure 3: Selection of accurate motion rhythm. The blue line shows the skeletal points' movement speed per frame, while the orange line represents the onset. The red and green dots indicate candidate motion rhythm points, with red dots indicating selected frames and green dots indicating unselected frames.

of our network. Our main approach involves feature extraction using multiple MFE (motion feature extraction) blocks. M2BNet has been trained using the AIST-M2B dataset.

Network Architecture. We propose a spatial-temporal graph model, M2BNet, to predict motion rhythm by incorporating temporal and spatial features. While 2D joint locations provide a natural representation of human motion, the skeletal structure is in the form of graphs rather than 2D or 3D grids, posing challenges for effective capture of spatial characteristics by CNNs. Building upon previous work [Aasen, 2021; Yu *et al.*, 2017; Liu *et al.*, 2020; Song *et al.*, 2020], our M2BNet using MFE block addresses this limitation and offers an improved approach for modeling human motion.

Each MFE block consists of a residual structure, a skeleton graph convolution operator, and a temporal convolution structure. The SGCN (skeleton graph convolution network) extracts features from a single frame in the spatial domain, and the TCN (temporal convolution network) processes temporal information across multiple frames.

1) SGCN. We use 2D keypoints as input where each value represents the position of a lamb. These keypoints are in a graph structure and not continuous features. Compared to the transformer model [He *et al.*, 2023], GCN has significant advantages in processing inputs with graph structures, so we propose SGCN for extracting features from sparse matrices.

Human motion can be represented as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of N body joints and $\mathcal{E}$ represents the set of bones. Let $\mathbf{Y} \in \mathbf{R}^{T\times N\times C}$ denote the input feature, where T, N, and C represent the temporal length, number of keypoints, and number of channels respectively. For an input motion feature $\mathbf{Y}_{in}$, the output feature $\mathbf{Y}_{out}$ is extracted through graph convolution. The graph convolution is computed as follows:

$$\mathbf{Y}_{out} = \sum_{k=1}^{K} \bar{\mathrm{A}}_k \mathbf{Y}_{in} W_k, \tag{4}$$

where K is the spatial partition of the human skeleton. $\mathrm{A}_k \in \{0,1\}^{N\times N}$ is an adjacent matrix that defines the adjacent relationship of joints. $\bar{\mathrm{A}}_k = \Lambda_k^{-\frac{1}{2}} A_k \Lambda_k^{-\frac{1}{2}}$ is the normalized adjacent matrix. $\Lambda_k^{ii} = \sum_j (A_k^{ij}) + \alpha$. Λ_k is a degree matrix used to normalize the adjacency matrix A_k, and α is set to prevent empty rows. $W_k \in \mathbf{R}^{C_l \times C_{l+1}}$ denotes a learnable weight matrix at layer l. As shown in Figure 2 (c), the right leg represents a spatial partition, with yellow, green, and red points indicating the centripetal node, root node, and centripetal node, respectively. By employing this partitioning strategy, the SGCN network can utilize the spatial motion information from a single frame effectively.

2) TCN. Motion can be represented as a sequence of continuous keypoints. In addition to using information from single frames, temporal information is also crucial. We introduce TCN to extract motion features in temporal dimension.

Following the approach of naturally connecting human joint points in spatial dimensions, the same keypoints can be linked through consecutive time steps. This enables us to easily extend from a single frame to a temporal sequence. The network architecture of TCN is shown in Figure 2 (c). M2BNet consists of 10 MFE blocks, the TCN has a kernel size of (9, 1) in all MFE blocks. In the 5th and 8th MFE blocks, the TCN has a stride of (2, 1), while the others have a stride of (1, 1). We use a residual design in all MEF blocks except the first one.

Loss Function. The goal of our network is to determine whether a frame can be classified as motion rhythm. A direct approach is to use binary cross-entropy loss (BCE Loss) for training. However, there are significantly fewer rhythmic frames than non-rhythmic frames, resulting in sample imbalance in the training data. To address this issue, we employ a weighted version of the BCE loss. The weighted BCE loss is defined as follows.

$$\mathcal{L}_{W\text{-}BCE} = -(1-w) * y \log(\hat{y}) - w * (1-y) \log(1-\hat{y}), \tag{5}$$

where $\hat{y}$ is the predicted result and y is the ground truth. The coefficient w is derived from the distribution of labels within the dataset. We set w to 30 based on the distribution of rhythmic and non-rhythmic labels in the AIST-M2B dataset.

The output of M2BNet is a probability ranging from 0 to 1. We consider frames with confidence scores exceeding 0.9 as representing rhythmic moments empirically. If there are consecutive frames with scores exceeding 0.9, the highest confidence one is selected as the detected rhythmic moment.

3.3 Motion Rhythm Enhancement

Manual video editing for rhythm enhancement is a laborious and time-consuming process. In this study, we propose an automated method that utilizes assessed motion rhythm to generate rhythm-enhanced videos. Considering human sensitivity to music frequencies, even slight accelerations can lead to dissonance. Therefore, we introduce modifications to the video content in order to synchronize it with the music rhythm. Our proposed method comprises two steps: first, aligning the motion-video rhythms; second, enhancing the overall rhythmic quality.

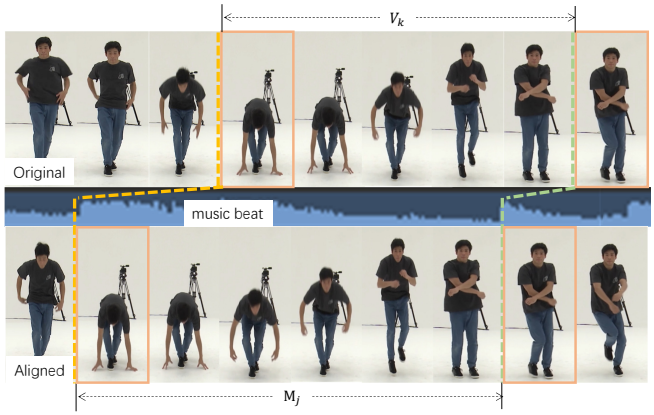

Figure 4: Rhythm Alignment. The motion is adjusted to match the music rhythm, and the durations V_k and M_j are recorded to calculate the score Dis_k and $offset_k$.

Rhythm Alignment. To ensure temporal alignment between motion and music rhythm, it is essential to calculate the optimal mapping rhythm. The methodology for achieving rhythm alignment is illustrated in Figure 4. The first row depicts the motion before rhythm alignment, while the second row represents the sound intensity waveform of the music. The third row displays the video after alignment. Here, V_k denotes the k-th motion beat, and M_j represents the corresponding index of a music beat for each motion beat (k). The time sequence of motion rhythm is denoted as V, with a length of $vlen$. Similarly, the time sequence of music rhythm is represented by V, having a length of $mlen$. Additionally, we define "Dis" as the time difference required for matching motion-video rhythms. For every motion rhythm (V_k), we can compute its minimum matching distance (Dis_k) with respect to a music rhythm (M_i).

$$Dis_k = \min_{i=0}^{mlen}\{|V_k - M_i + offset_{k-1}|\}, \tag{6}$$

where k represents the motion rhythm index, i represents the music rhythm index, and *offset* is the cumulative offset of all previously matched rhythms. It is necessary to add each offset, which is equal to $V_k - M_j$, to account for the difference between them. The formula of *offset* is as follows:

$$offset_k = \sum_{i=0}^{k}(V_i - M_i), \tag{7}$$

where j represents the music rhythm index corresponding to the k-th motion rhythm.

Rhythm Enhancement. In our approach, we employ a widely used motion video editing technique to enhance the overall rhythm of motion. The entire sequence can be segmented based on the motion rhythm, with each segment further divided into two parts: an acceleration part and a deceleration part. The deceleration part incorporates motions synchronized with the music rhythm, while the acceleration part facilitates smooth transitions between consecutive motions. This methodology enables viewers to perceive the intricate relationship between motion and music rhythm.

4 Experiment

4.1 Experimental Settings

Dataset Settings. AIST-M2B, derived from AIST++, is developed to acquire motion rhythm labels of superior quality by exploiting the inherent correlation between dance and music in the videos. It stands as the pioneering motion rhythm dataset, while diverse methods are employed for motion rhythm detection on this dataset.

Benchmark Models and Training Protocols. In order to assess the performance of M2BNet on AIST-M2B, a random selection is made where 90% of the videos are assigned as the training set, while the remaining 10% serves as the validation set. The training process encompasses 200 epochs with a learning rate of 0.001. The model that exhibits superior performance on the validation set is chosen for testing purposes, with a confidence threshold set to 0.9.

MuDaR [Yu *et al.*, 2023] consider music rhythm and motion rhythm to be equivalent. Since the latter is not available in open source, we use music rhythm as the training set, referred to as M2B(music).

Evaluation Metrics. Motion rhythm prediction can be regarded as a series of binary classification tasks, where each frame is assigned a probability label. Considering the temporal continuity of motion rhythm, predictions within a certain time window are considered accurate if their difference from the ground truth value falls below a threshold. In this study, we set the time window to 0.07 seconds.

In music rhythm evaluation, F-measure and Cemgil are used to assess the similarity between predicted beats and dataset's beats. The F-measure evaluates the accuracy of detected beats by marking them when they fall within a specific range around the dataset's beats. Cemgil measures the precision of detected beats, with higher scores given to closer beats in relation to dataset's beats. These metrics used with default parameters to evaluate the similarity of motion beat labels.

4.2 Results and Analysis

Motion Rhythm Extraction. We test the results of both traditional methods and learning-based methods. The experimental results are presented in Table 1.

Method	Accuracy	Recall
[Davis and Agrawala, 2018]	0.32	0.98
[Bellini *et al.*, 2018]	0.24	0.96
[Chu and Tsai, 2011]	0.35	0.91
M2B(music)	0.36	0.39
M2B	**0.60**	0.59

Table 1: Precision and recall of the motion rhythm extraction methods. M2B(music) notes our model trained with music rhythm labels. The first three lines represent mathematical methods, the last two lines represent deep learning methods.

The visualization results of M2B are displayed in Figure 5. The upper half represents the video input, while the lower

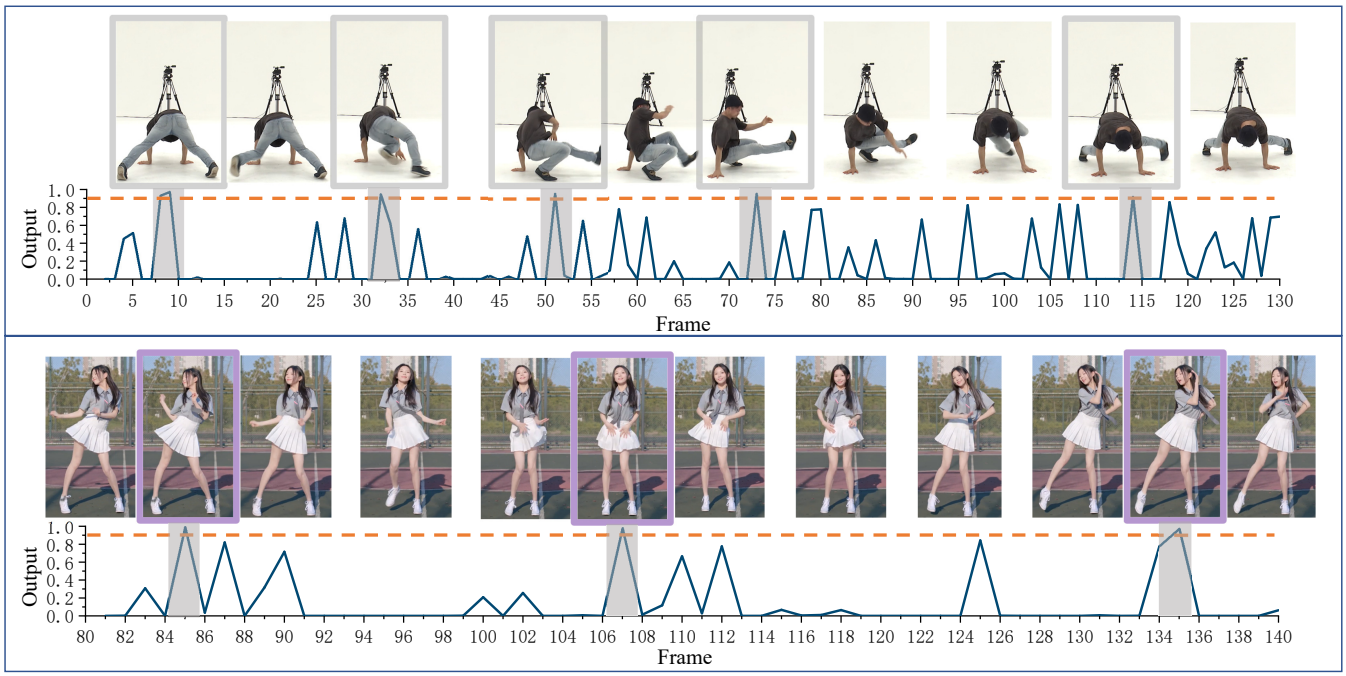

Figure 5: Visualization of model output. The top half shows the video images, the bottom half is the model's output. The orange line represents the confidence score. Values exceeding the confidence threshold are selected as motion rhythms (images with boxes). We present images around the motion rhythm for comparison. The displayed video is sourced from *AIST++* and *BiliBili*.

half depicts the model output. Confidence is indicated by the red line, and values exceeding the confidence threshold are selected as motion rhythm.

Dataset Analysis. To validate the relevance of each dataset, we invite ten experts in the field of dance rate the datasets. We randomly select twenty videos from each dataset and request experts annotate motion rhythm, used F-measure and Cemgil scores to evaluate the similarity between expert annotations and dataset. M2B (music) annotated with music rhythm labels get F-measure of 0.54, Cemgil of 0.39. Our AIST-M2B get F-measure of **0.65**, Cemgil of **0.44**.

The experiments demonstrate that our dataset aligns better with human perception of motion rhythm. Please refer to the supplementary materials for the visualization charts of the dataset.

Motion Enhancement. In motion enhancement, we calculate the optimal mapping relationship for each motion beat and use Cemgil to evaluate its effectiveness. Videos and music are randomly selected from the dataset. We test the performance of different methods in enhancing rhythm. Based on the editing process, we can infer the enhanced rhythm from the original rhythm. The raw videos achieve a Cemgil score of 0.53, compared to 0.63 for visbeat [Davis and Agrawala, 2018] and **0.82** for our M2BNet.

4.3 Ablation Study

The effectiveness of each module in M2BNet is illustrated through an ablation study conducted on AIST-M2B. M2BNet comprises multiple SGCN, TCN, and residual connections, with one-dimensional features being extracted by the FC-layer. Separate tests are performed on M2BNet after reducing the SGCN, TCN, RES, and FC modules. The experimental results are presented in Table 2.

	Precision	Recall
w/o SGCN	0.49	0.53
w/o TCN	0.28	0.42
w/o RES	0.59	0.58
w/o FC	0.57	0.58
M2B	**0.60**	**0.59**

Table 2: Ablation study on the M2BNet structures.

5 Conclusion

In this study, we present AIST-M2B, the first motion rhythm dataset with automatic high-quality annotations, to our knowledge. We propose M2BNet, a novel approach that leverages spatial and temporal information to assess motion rhythm. Additionally, we introduce a method for enhancing motion rhythm to reduce video editing costs. However, our method can be further improved in scenarios where body parts go out of the camera's frame. For future research, we will further enhance the accuracy and generalizability of motion rhythm assessment, introduce M2Beats 2.0, and make more short-form videos worthwhile.

Acknowledgements

This work was supported by the Innovation Research Group Project of NSFC (61921003).

Contribution Statement

Dongxiang Jiang and Yongchang Zhang have equally contributed (denoted by † on Page 1) to conceptualization, methodology, investigation, formal analysis and writing. Shuai He has provided resources and visualization. Anlong Ming, the corresponding author, has contributed to conceptualization, methodology, funding acquisition, resources, supervision, as well as review and editing.

References

[Aasen, 2021] Solveig Aasen. Crossmodal aesthetics: How music and dance can match. *The Philosophical Quarterly*, 71(2):223–240, 2021.

[Bellini *et al.*, 2018] Rachele Bellini, Yanir Kleiman, and Daniel Cohen-Or. Dance to the beat: Synchronizing motion to audio. *Computational Visual Media*, 4:197–208, 2018.

[Böck and Widmer, 2013] Sebastian Böck and Gerhard Widmer. Maximum filter vibrato suppression for onset detection. In *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, volume 7, page 4, 2013.

[Böck *et al.*, 2016] Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. Madmom: A new python audio and music signal processing library. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1174–1178, 2016.

[Bredin and Chollet, 2007] Hervé Bredin and Gérard Chollet. Audiovisual speech synchrony measure: application to biometrics. *EURASIP Journal on Advances in Signal Processing*, 2007:1–11, 2007.

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

[Chu and Tsai, 2011] Wei-Ta Chu and Shang-Yin Tsai. Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos. *IEEE Transactions on Multimedia*, 14(1):129–141, 2011.

[Chung and Zisserman, 2017] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part II 13*, pages 251–263. Springer, 2017.

[Davis and Agrawala, 2018] Abe Davis and Maneesh Agrawala. Visual rhythm and beat. *ACM Transactions on Graphics (TOG)*, 37(4):1–11, 2018.

[Ge *et al.*, 2021] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021.

[Halperin *et al.*, 2019] Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. Dynamic temporal alignment of speech to lips. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3980–3984. IEEE, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2022] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *IJCAI*, pages 942–948, 2022.

[He *et al.*, 2023] Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1023–1032, 2023.

[Kim *et al.*, 2003] Tae-hoon Kim, Sang Il Park, and Sung Yong Shin. Rhythmic-motion synthesis based on motion-beat analysis. *ACM Transactions on Graphics (TOG)*, 22(3):392–401, 2003.

[Liu *et al.*, 2020] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang. Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152, 2020.

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

[Owens and Efros, 2018] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018.

[Sargin *et al.*, 2007] Mehmet Emre Sargin, Yücel Yemez, Engin Erzin, and A Murat Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE transactions on Multimedia*, 9(7):1396–1403, 2007.

[Shiratori and Ikeuchi, 2008] Takaaki Shiratori and Katsushi Ikeuchi. Synthesis of dance performance based on analyses of human motion and music. *Information and Media Technologies*, 3(4):834–847, 2008.

[Shiratori *et al.*, 2004] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Detecting dance motion structure through music analysis. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 857–862. IEEE, 2004.

[Shiratori *et al.*, 2006a] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Dancing-to-music character animation. In *Computer Graphics Forum*, volume 25, pages 449–458. Wiley Online Library, 2006.

[Shiratori *et al.*, 2006b] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Dancing-to-music

character animation. In *Computer Graphics Forum*, volume 25, pages 449–458. Wiley Online Library, 2006.

[Song *et al.*, 2020] Chao Song, Youfang Lin, Shengnan Guo, and Huaiyu Wan. Spatial-temporal synchronous graph convolutional networks: A new framework for spatial-temporal network data forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 914–921, 2020.

[Tchemeube *et al.*, 2023] Renaud Bougueng Tchemeube, Jeffrey Ens, Cale Plut, Philippe Pasquier, Maryam Safi, Yvan Grabit, and Jean-Baptiste Rolland. Evaluating human-ai interaction via usability, user experience and acceptance measures for mmm-c: A creative ai system for music composition. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 5769–5778, 2023.

[Wang *et al.*, 2020] Jianren Wang, Zhaoyuan Fang, and Hang Zhao. Alignnet: A unifying approach to audio-visual alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3309–3317, 2020.

[Yu *et al.*, 2017] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875*, 2017.

[Yu *et al.*, 2023] Jiashuo Yu, Junfu Pu, Ying Cheng, Rui Feng, and Ying Shan. Learning music-dance representations through explicit-implicit rhythm synchronization. *IEEE Transactions on Multimedia*, pages 1–10, 2023.

[Zhou *et al.*, 2023] Qiu Zhou, Manyi Li, Qiong Zeng, Andreas Aristidou, Xiaojing Zhang, Lin Chen, and Changhe Tu. Let's all dance: Enhancing amateur dance motions. *Computational Visual Media*, 9(3):531–550, 2023.