UNIVERSAL PHYSICAL EDUCATION: TRAINING-FREE ACTION QUALITY ASSESSMENT BASED ON DYNAMIC TIME WARPING

Anonymous authors
Paper under double-blind review

ABSTRACT

The increasing integration of video-based evaluation into sports education underscores the need for efficient and adaptable methods to assess action quality. Existing approaches, often reliant on deep learning models and extensive labeled datasets, face challenges in generalizing to novel exercises or scenarios with limited training data. This paper introduces the Multi-Dimensional Exercise Distance Adaptive Constrained Dynamic Time Warping (MED-ACDTW) method, a training-free algorithm designed for action quality assessment. By leveraging features extracted from both 2D and 3D spatial dimensions and integrating multiple human body attributes, our method achieves a 2-3% improvement in accuracy over single-dimensional pose estimation techniques. The proposed adaptive constraint scheme further enhances action quality discriminability, achieving a 30% improvement in precision compared to conventional methods. Additionally, we introduce BGym, a novel dataset tailored for evaluating action quality in diverse and non-standardized perspectives. The MED-ACDTW method eliminates the dependency on labeled datasets, enabling direct comparison between template and test videos, thus offering a scalable and versatile solution for real-world applications.

1 INTRODUCTION

With the advancements in online sports instruction, the extensive volume of video homework submissions often overwhelms physical education instructors, inhibiting timely feedback and objective assessment. Automated action quality assessment scoring systems can address these issues by providing accurate and real-time evaluation.

One approach to action quality assessment is via action matching. Effenberg et al. (2016) collected data related to rowing motion parameters, utilized the DTW algorithm to calculate the distance matrix, and then combined it with the force index to construct a universal distance value to match with actions. Manousaki et al. employed a variant of DTW to predict action labels by comparing the time series of unfinished actions with those of known prototype action executions. DP-DTW (Chang et al., 2021) learns a discriminative prototype for each category and distinguishes different motion categories by computing the DTW outputs between the input sequence and prototypes of different categories. However, a crucial problem they will face is that applying DTW and ED distances to the scoring problem leads to a significantly low level of discriminability. Although actions are distinguished, the score differences among different actions are not substantial.

Another method for assessing action quality is regression-based approaches. The main challenge of regression methods comes from the need for labeled data, resulting in AQA datasets usually being of small scale. Parmar et al. proposed the MTL-AQA (Parmar & Morris, 2019b) dataset, which contains only 1,000 samples. In the same year, they proposed AQA-7(Parmar & Morris, 2019a), which also has only 500 samples. Xu et al. (2022) proposed FineDiving, which has only 3,000 samples. Subsequently, many deep learning regression models (Tang et al., 2020; Zhang et al., 2023; Zhou et al., 2023; Farabi et al., 2022) have been applied on these datasets. However, acquiring various labels requires professional assessment work, which goes against our original intention of

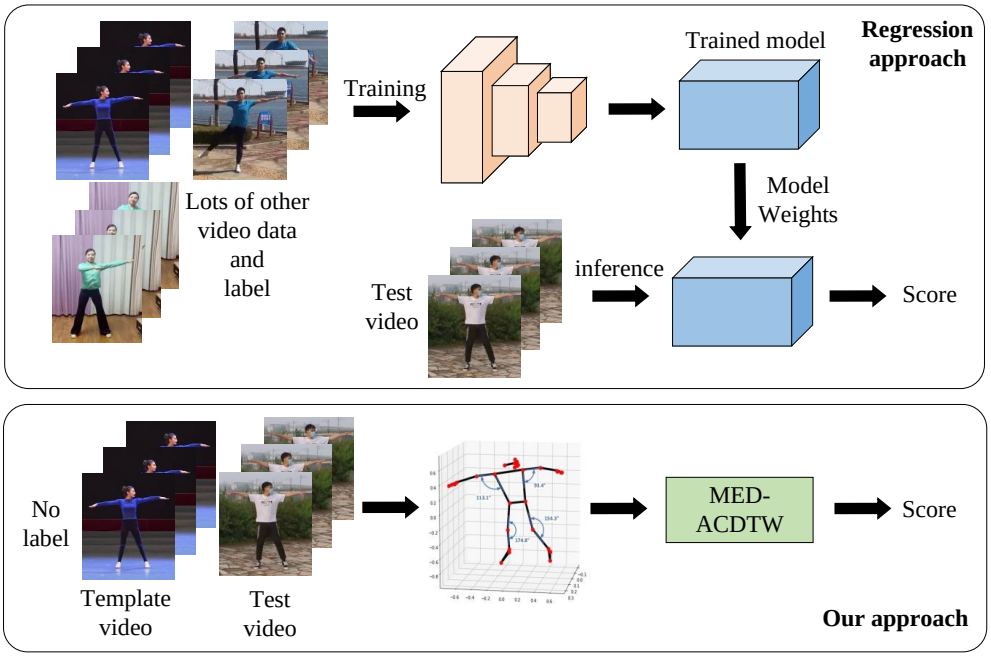

Figure 1: Two methods of action quality assessment. The above part represents the regression method of deep learning. The bottom part represents the action matching method for our works.

automation. When migrating to other action-type datasets, fine-tuning the pre-trained model on small-scale AQA datasets often has a serious overfitting risk.

In addition, some studies have incorporated additional text labels to enhance action quality assessment (Du et al., 2023; Zhang et al., 2024; Zeng & Zheng, 2024; Majeedi et al., 2025). The primary aim of introducing these text labels is to address the limited distinguishability of visual features in small datasets while bolstering the model's ability to comprehend fine-grained action quality. However, this method fundamentally succumbs to the "label dependency trap," which directly contradicts the core goal of automated evaluation. It demands more specialized and granular annotation content. For instance, to enable large language models (LLMs) to align textual descriptions with visual action segments, researchers must annotate each sample with multidimensional textual details. In diving datasets, this might include specifics such as "deviation of entry angle from vertical" and "size of entry splash (small/medium/large)." Such processes require domain experts to invest 5-10 minutes per sample in a detailed description and validation.

To overcome these issues, this paper introduces a novel action quality assessment algorithm MED-ACDTW. We observe that in sports settings such as schools and gyms, instructors, coaches, and online fitness platforms often provide standard template videos for beginners to learn from. Building on this, we introduce a comparative approach that evaluates the representation of test videos by comparing them to these template videos. This indirect evaluation of action quality reduces the dependence on labels. It allows for the utilization of template videos as an alternative label attribute, offering greater flexibility for real-world applications. Figure 1 shows the difference between our scheme and the supervised learning regression method. Regression methods require a large number of videos and labels for training to save the model weights with the best training effect. During the inference process, videos of the same type as those in the training set can be input for scoring. An additional drawback is that the length of the videos in the training set will limit the length of the test video. In contrast, our solution only requires inputting a standard video and a test video. The score can then be obtained through skeletal extraction and the MED-ACDTW algorithm, which better reflects the end-to-end advantage. Moreover, in practical sports applications, we can handle videos of any action.

Secondly, to address the issue of inadequate accuracy in DTW-based evaluation of time series, this paper proposes two solutions for feature and frame distance calculation. Regarding features, the focus is on joint angle features, directional features, and upper-lower limb coordination features. To handle the distance calculation issue arising from these features, the paper proposes the MED module, which employs a ratio-based approach to measure the impact of each feature. Thirdly, classic DTW algorithms are not robust to outliers, as the ideal DTW path should follow a diagonal pattern, while outliers can distort the path. This distortion leads to excessive stretching and compression (Wan et al., 2017). To tackle this issue, adaptive constraints are applied in the context of action quality assessment. Specifically, we make the following contributions:

(1) **Training-Free Action Matching Framework**: We propose the MED-ACDTW algorithm, a novel training-free approach that directly compares template videos with test videos, eliminating the need for laborious annotation processes such as action difficulty scoring. This algorithm offers significant flexibility and is applicable across a wide range of video comparison tasks beyond physical education.

(2) **Adaptive Constraint Mechanism**: To address the common limitations of traditional DTW algorithms, such as sensitivity to outliers and misaligned frame mappings, we introduce an adaptive constraint scheme. This innovation ensures robust and precise alignment of action sequences by dynamically adjusting constraints, representing the first successful application of this technique in action quality assessment.

(3) **Comprehensive Feature Set with Enhanced Discriminative Power**: We design a multi-dimensional feature set that integrates joint angles, directional movements, and limb coordination from both 2D and 3D skeletal data. These features enable a fine-grained representation of human body pose. To apply these features to DTW, we introduce the MED (Multi-Dimensional Exercise Distance) method, which surpasses traditional Euclidean distance in its ability to capture subtle differences in body movement. The MED method significantly enhances the discriminative power of action matching, offering improved accuracy in distinguishing between similar actions.

(4) **BGym Dataset Creation**: We construct BGym, a standardized dataset tailored for action quality assessment in physical education. Consisting of 153 high-quality, cropped video samples with corresponding evaluation labels, this dataset fills a critical gap in the field by providing a benchmark for developing and validating action scoring algorithms. It is expected to facilitate future research and application development in this domain.

2 Related Work

2.1 Action Quality Assessment in Regression-based approach

Pirsiavash et al. (2014) was one of the first to introduce deep learning to assess movement quality in sports, using regression models based on spatiotemporal pose features. Parmar & Morris (2019a) proposed learning a single model across multiple actions, which utilizes the C3D model for feature extraction. Farabi et al. (2022) employed the ResNet network for feature extraction and utilized the Weight-Decider technique to handle downstream scoring tasks, demonstrating its capability to handle 32-frame segments. Tang et al. (2020) introduced the I3D-MLP scheme, which leverages the I3D model for extracting features. They further implemented a scoring distribution method to accurately represent the probability of action quality assessment scores, yielding precise outcomes. Building upon Tang et al.'s foundation, Zhang et al. (2023) introduced the I3D-DAE scheme. This innovative approach includes a plug-and-play module called the Distribution Auto-Encoder (DAE), which enhances the accuracy of mapping videos to scores. Zhou et al. (2023) introduced a sophisticated hierarchical graph convolutional network (HGCN), designed to address the challenges that arise when video sequences are divided into uniformly sized clips. This division often results in confusion within clips and incoherence between them. Their approach significantly improves the discriminative features of the videos, thereby elevating the performance of action quality assessment.

Zeng & Zheng (2024) additionally adopts optical flow and audio beyond sole visual input to propose the Progressive Adaptive Multimodal Fusion Network (PAMFN), which includes three modality-specific branches for independent feature modeling, a mixed-modality branch for progressive ag-

gregation. Zhang et al. (2024) investigates a new task called narrative action evaluation (NAE) and proposes a prompt-guided multimodal interaction framework—its key features include using a pair of transformers to facilitate multimodal information interaction and converting the score regression task into a video-text matching task to enable task interactivity. Majeedi et al. (2025) propose RICA², a deep probabilistic model for action quality assessment (AQA) that additionally integrates a score rubric and accounts for prediction uncertainty, featuring stochastic embeddings of action steps on a graph structure encoding the score rubric to spread probabilistic density and represent model uncertainty. Xu et al. (2025) targets the long-term sports assessment task in the long video domain and proposes the language-guided audio-visual learning (MLAVL) framework, which models "audio-action-visual" correlations under the guidance of low-cost language modality and integrates components like action knowledge graphs, a shared-specific context encoder, an audio-visual cross-modal fusion module, and a dual-branch prompt-guided grading module.

2.2 Skeleton Data Acquisition

In the field of skeleton information acquisition, researchers have explored multiple modalities. First of all, MediaPipe Lugaresi et al. (2019), a lightweight pose estimation model, has demonstrated high accuracy and low latency pose detection, making it well-suited for mobile devices. Zhang et al. (2020) utilized MediaPipe in a framework for developing cross-platform machine learning solutions. Their proposed model and pipeline architecture exhibited real-time inference speeds on mobile GPUs and high prediction quality. MediaPipe also finds applications in medical analysis. For instance, Güney et al. (2022) employed MediaPipe to quantitatively evaluate hand movements in patients with Parkinson's disease, proposing automatic tremor analysis.

2.3 Feature Information Representation

After processing the skeleton data, the derived features can capture various postures and reflect the nuances of human movement. Hu et al. (2015) constructed pose features based on the relative positions of skeletons, considering the spatiotemporal trajectory of their movements. Rahmani et al. (2014) utilized the displacement information of skeletons to compute the maximum sweep area, which represents the postural characteristics of the human body. The directional features of joint limbs can encode joint angle information. Gu et al. (2012) calculated the direction of the displacement vector using 15 skeletons, extracting joint angle features that represent the human torso and reflect the relative direction of the joints. Bakhat et al. (2023) observed the joint order of 3D joints, extracted temporal information, and utilized zero-crossing features between current and previous frames to derive reliable features. They proposed an OMKZ scheme using ordered, modified Katz centrality and zero-cross information for human motion recognition.

3 Our Approach

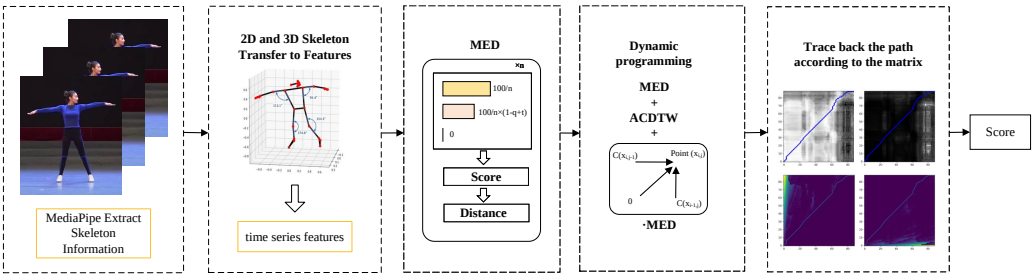

Figure 2: The overall structure and workflow of the MED-ACDTW.

In this section, we first introduce how to extract the features of skeleton joint points. Then, we describe the MED-ACDTW method in detail. The architecture of MED-ACDTW for action quality assessment is shown in Figure 2. Modules 1 and 2 show the process of feature extraction and manual feature construction. The process of extracting continuous frames of 2D and 3D key points

using MediaPipe. After that, the extracted keypoints enter the human body features building block, which includes four types of features. Modules 3, 4 and 5 represent our action matching algorithm. The constructed frame-based time series features are entered into the third module, and the MED algorithm is used to calculate interframe fractions and distances. Then the calculated distance is input into ACDTW module, and adaptive constraints are introduced to obtain the constraint matrix and score matrix. Backtracking is performed to calculate the total score based on score matrix.

3.1 FEATURE EXTRACTION

When relying solely on skeleton point coordinates for DTW, the accuracy can be insufficient due to individual variations in body structure and movements. To overcome this limitation, we propose constructing both local features, which capture the relationships between different body joints, and global features, which reflect overall movement coordination. The detailed construction of these features is outlined in Table 1.

Table 1: All feature details

Type	Name	Num
Barycentric	Human Body Angle	1
Joint Angle	Limb Mass Center Point	4
	Limb Angle	13
Direction	Human Body 2D Direction	2
	Pelvic Horizontal Angle, Pelvic Rotation Angle	2
	Shoulder Horiz. Angle, Shoulder Rotation Angle	2
Upper and Lower Limb Linkage	Forearm & Calf Angle	2
	Forearm & Thigh Angle	2
	Foot L. & Shoulder W. Ratio	1
	Wrist L. & Shoulder W. Ratio	1

3.1.1 JOINT ANGLE FEATURE

For feature construction, joint points are shown in Figure 3(a), where nodes 0 to 32 represent default nodes in MediaPipe. In the angular feature construction, we build two features. First, we aim to establish connections between adjacent joints in the human body. To achieve this, we calculate the angles between 13 adjacent joints, focusing on key areas such as elbows, shoulders, and knees. As shown in 3(b). The second involves characterizing the deviation of limb block center points. We consider each of the four limbs as a separate region, depicted by a polygon formed by specific joint points within that region. We then compute the center point of each region and calculate the Angle between that point and the spine vector. The calculation method is shown in Eq.equation 1 and Eq.equation 2, where p_0 represents the coordinates of the center point. x_i, y_i, z_i respectively represent the vertices of the polygon. $\overrightarrow{M}$ stands for spine vector.

$$p_0 = \left(\frac{\sum_{i=1}^{4} x_i}{4}, \frac{\sum_{i=1}^{4} y_i}{4}, \frac{\sum_{i=1}^{4} z_i}{4} \right) \tag{1}$$

$$\theta = \arccos \left(\frac{\overrightarrow{M} \cdot \overrightarrow{op_0}}{|\overrightarrow{M}| \times |\overrightarrow{op_0}|} \right) \tag{2}$$

3.1.2 BARYCENTRIC FEATURE

In order to determine the barycentric feature, we employed the coefficient method (Ying, 2021). The process implies multiplying the coordinates of each joint point by the corresponding inertia parameters of the human body to yield the body's center of gravity or barycenter. Also, calculate the angle between the Barycentric and the spine to form a feature.

3.1.3 DIRECTION FEATURE

The direction features are constructed using the angles of the hip and shoulders, as shown in Figure 3(c) and (d). To construct the features for the pelvic horizontal angle and the pelvic anterior tilt angle, we calculate the angle between the offset direction of these body parts and the coordinate axes. The same methodology is applied to the shoulder angles, including the shoulder horizontal and anterior tilt angles.

Table 1 presents a comprehensive breakdown of the feature details, consisting of a total of 30 distinct features. We compute both the central and angular features simultaneously in both 2D and 3D dimensions. As a result, the total count of features is 52, comprising 30 features in 2D and 22 features in 3D.

3.1.4 UPPER AND LOWER LIMB LINKAGE FEATURE

While the previously mentioned features mainly focus on the upper or lower limbs individually, there is a need for features that effectively bridge the upper and lower limbs. To address this, we introduce new features based on the angles between the forearm and calf, and the forearm and thigh. These additional features enhance our ability to evaluate the linkage between the upper and lower limbs during actions such as standing upright or sitting. For the features that involve the forearm and calf, using the left side of the body as an example, $\overrightarrow{AB}$ represents the left forearm, $\overrightarrow{CD}$ represents the left calf. The linkage features of the left half of the body can be calculated using Eq.equation 2. Similarly, calculations are performed for the calf and thigh.

Additionally, we utilize the ratio of the tensioned feet to the shoulder width and the ratio of the tensioned hands to the shoulder width to assess the tension in the limbs. These features are effective in distinguishing between upright and stretching motions. To appropriately integrate these features into the human posture features framework of this study, it is necessary to convert the non-radian numbers into radians. Calculate the 3D distance between both ankle joints and the 3D distance between shoulder joints. If the distance between the ankles is greater than half the distance between the shoulders, then $\theta_1 = 3\pi/2$; if it is less, $\theta_1 = \pi/2$. This provides the feature θ_1 - the ratio of the feet's tension degree to the shoulder width. The same design is used to calculate the 3D distance between the wrist and shoulder joints.

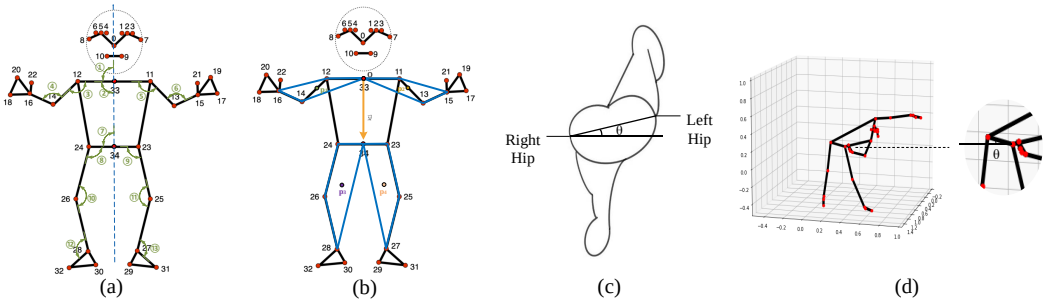

Figure 3: Feature presentation. (a) represents the joint points of humans identified by MediaPipe. (b) represents the Pelvic Horizontal Angle. (c) represents the Pelvic Rotation Angle.

3.2 ACTION MATCHING ALGORITHM

The Action matching algorithm incorporates the MED module to transform multi-dimensional features into feature score matrices and calculate inter-frame motion distances. Furthermore, a dy-

namic programming matrix is generated using the concept of DTW, with the addition of adaptive constraints to penalize deviations from the path. Finally, the optimal solution is found using the backtracking method based on the distance matrix of the dynamic programming matrix, which is then mapped to the score matrix. The final score is obtained by accumulating the scores and computing the average.

3.2.1 MED MODULE

The MED module is utilized to calculate the score between two frames. The following definitions are used:

- n,m represent the number of frames in the template video and the test video respectively, and h represents the number of features.
- $X = \{x_1, x_2, \cdots, x_n\}$ and $Y = \{y_1, y_2, \cdots, y_n\}$ represent the values of each feature in a test video frame and a template video frame, respectively.
- l is the length of the scoring path in the 2D matrix.
- $Score_{1,1}$, $Score_{1,2}, \cdots$, $Score_{\alpha,\beta}$ are the scores calculated for two frames. Each pair of frames represents a single point on the two-dimensional matrix.
- FS denotes the final score measured for the two videos.

To calculate the score, each feature score is computed individually. As shown in Eq.equation 3, the sum of all feature scores totals 100 points. Next, the corresponding limb features of the template and the test video are subtracted to determine their difference. The ratio of this difference to the feature of the template video is then calculated. Based on the value of this ratio, denoted as q, the score for each feature is determined. According to Eq.equation 4 and Eq.equation 5, if q is less than the threshold parameter t, it is considered an acceptable error, and the score remains unchanged.

$$S_1 = S_2 = \ldots = S_n = \frac{100}{h} \tag{3}$$

$$q = \frac{|x_i - y_i|}{x_i} \tag{4}$$

$$S_i = \begin{cases} S_i, & q \leq t \\ S_i \times (1 - q + t), & t < q \leq 1 \\ 0, & q > 1 \end{cases} \tag{5}$$

The score for all features of two frames is calculated according to Eq.equation 6. The scores of all features are then summed to obtain the score between the two frames. Based on this score and Eq.equation 7, the distance between the two frames can be calculated, which replaces the Euclidean distance.

$$Score_{i,j} = \sum_{k=1}^{n} S_k \tag{6}$$

$$MED(i,j) = \frac{1}{Score_{i,j}} \times 100 - 1 \tag{7}$$

The score corresponding to each frame on the path is calculated using the ACDTW algorithm and the MED module. The sum of the scores on the path is then averaged to obtain the final score between the test video and the template video. This represents the score on the score path. The calculation of the path score is shown in Eq.equation 8. Algorithm 1 for calculating feature distance instead of Euclidean distance is shown below.

$$FS = \frac{\sum_{i=0,j=0}^{n,n} Score_{i,j}}{l} \tag{8}$$

Algorithm 1: Calculation of inter-frame distance via MED module

Input: Template video feature $x_i(i = 1 \dots n)$, Test video feature $y_i(i = 1 \dots n)$
Output: med

```
score = 0, n = len(x_i)
for i = 1 : n do
    t = 0.2
    q = abs(x_i − y_i)/x_i
    if q ≤ t then
        point_score = 100/n
    end
    else if q > 1 then
        point_score = 0
    end
    else
        point_score = 100/n × (1 − q + t)
    end
    score = score + point_score
end
med = score/100 − 1
return med
```

3.2.2 Adaptive Constrained DTW

For convenience, this paper initially provides the formula for the unrefined DTW, as shown in Eq.equation 9, to enable a comparison with the enhanced formula within the ACDTW. ED represents the Euclidean distance between features between two frames.

$$\begin{aligned} &DTW(1,1) = ED(1,1), \\ &DTW(i,j) = ED(i,j) + \min \begin{cases} DTW(i-1,j-1) \\ DTW(i-1,j) \\ DTW(i,j-1) \end{cases} \end{aligned} \tag{9}$$

Transitioning from the foundational DTW approach, the ACDTW method introduces a novel adaptive penalty function designed to address the specific challenge of multiple frames mapping to a single frame within an action time series. This adaptive approach seeks to optimize the usage of each frame in the action sequence by expanding the penalty function as the number of uses increases (Li et al., 2020). We employ a adaptive penalty function that takes into account varying trajectory lengths and strengthens the penalty function via dual constraints. Eq.equation 10 shows the adaptive penalty function.

$$C(x_{i,j}) = \frac{2\max(a,b)}{(a+b)} N(x_{i,j}) \tag{10}$$

In this penalty function, a and b represent the lengths of the two action time series respectively. $N(x_{i,j})$ denotes the frequency of each point's participation in the matching process. Meanwhile, the penalty exhibits proportional growth corresponding to an increasing magnitude of difference between a and b. This is because when two video frames significantly differ in length, they are likely not from the same category of video. As such, a larger penalty is applied to distinguish them. Thus, the value of the penalty function is directly proportional to the distance between the two time series, indicating a decreasing tolerance for many-to-one and one-to-many matches as the divergence between the two time series increases.

When the coefficient is zero, it corresponds to the standard DTW algorithm. Moreover, when the constraint coefficient is constant, $C(x_{i,j})$ is directly proportional to $N(x_{i,j})$. The computation of $N(x_{i,j})$ relies on two matrices, $T_P = (P_{i,j})\, m \times n$ and $T_Q = (Q_{i,j})\, m \times n$, which denote the frequency of participation in the matching process for P and Q respectively. The action sequence for frame p_i participates $P_{i,j}$ times, and the action sequence for frame q_i participates $Q_{i,j}$ times.

The sum of these two participation counts yields the value of $N(x_{i,j})$. To prevent the occurrence of multiple frames corresponding to a single frame or a single frame corresponding to multiple frames.

The state transition equation of the ACDTW is expressed in Eq.equation 11. In this penalty function, a and b represent the lengths of the two action time series respectively. $N(x_{i,j})$ denotes the frequency of each point's participation in the matching process. Meanwhile, the penalty exhibits proportional growth corresponding to an increasing magnitude of difference between a and b. This is because when two video frames significantly differ in length, they are likely not from the same category of video. As such, a larger penalty is applied to distinguish them. Thus, the value of the penalty function is directly proportional to the distance between the two time series, indicating a decreasing tolerance for many-to-one and one-to-many matches as the divergence between the two time series increases.

$$\begin{aligned} &ACDTW(1,1) = MED(1,1), \\ &ACDTW(i,j) = MED(i,j) + \\ &\min \begin{cases} ACDTW(i-1,j-1) \\ ACDTW(i-1,j) + C(x_{i-1,j})MED(i,j) \\ ACDTW(i,j-1) + C(x_{i,j-1})MED(i,j) \end{cases} \end{aligned} \tag{11}$$

If the trajectory from $(i-1, j-1)$ to (i, j) visits unique indices, the penalty function is not employed in this step. If the path from $(i-1, j)$ to (i, j) indicates the q_j frame is reused, the MED incorporates a penalty function, setting $P_{i,j} = 1$ and $Q_{i,j} = 1 + Q_{i-1,j}$. If the path from $(i, j-1)$ to (i, j) suggests that the p_i frame is reused, the MED similarly adds a penalty function, setting $Q_{i,j} = 1$ and $P_{i,j} = 1 + P_{i-1,j}$. This approach allows the MED-ACDTW algorithm to handle multi-dimensional action time series while preventing multiple frames from corresponding to a single frame. The pseudo-code summary of the ACDTW algorithm is shown in Algorithm 2.

Algorithm 2: ACDTW algorithm

Input: Template video feature $x_i(i = 1 \dots n)$, Test video feature $y_i(i = 1 \dots n)$
Output: Optimal warping path, distance between x and y

```
Initialize: P_{1,1} = 1, P_{i,j} = 0, i = 2, ··· , m, j = 2, ··· , n.
  Q_{1,1} = 1, Q_{i,j} = 0, i = 2, ··· , m, j = 2, ··· , n. ACDTW(1,1) = MED(1,1) ;
for i = 1 : m do
    for j = 1 : n do
        D1 = MED(i,j) + ACDTW(i-1,j-1)
        D2 = MED(i,j) + c(x_{i-1,j})MED(i,j) + ACDTW(i-1,j)
        D3 = MED(i,j) + c(x_{i,j-1})MED(i,j) + ACDTW(i,j-1)
        ACDTW(i,j) = min(D1, D2, D3)
        Optimal path[i,j] = min_index((i-1,j),(i-1,j-1),(i,j-1))
        if ACDTW(i,j) == D1 then
            P_{i,j} = 1, Q_{i,j} = 1;
        end
        else if ACDTW(i,j) == D2 then
            P_{i,j} = 1, Q_{i,j} = 1 + Q_{i-1,j};
        end
        else
            P_{i,j} = 1 + P_{i,j-1}, Q_{i,j} = 1;
        end
    end
end
return path, P, Q;
```

4 EXPERIMENTS

In this section, we first introduce the metrics and datasets used in this article. Next, we apply the MED-ACDTW algorithm to the constructed dataset, optimizing parameters through ablation exper-

iments. Then, we compare the performance of our algorithm against other mainstream methods. Finally, we present and interpret the results through visualization.

4.1 EXPERIMENTAL SETUP

The experimental setup involves matching an action with the highest score and comparing it against the true action for consistency. A match is successful if the same type of action achieves the highest score, implying that all scores of different action types are lower. Conversely, if a different action type achieves the highest score, it is considered a mismatch. Consequently, the matching success rate (accuracy) serves as an evaluation metric for our approach, as shown in Eq.equation 12.

Furthermore, to assess action quality, similar actions aim to score above 80, while different types of actions should score below 80. This distinction aids in applying the methods proposed in this study to practical scenarios. Therefore, we introduce an evaluation metric, $Rate80$, to indicate this discrimination effect. Eq.equation 13 represents the total number of similar actions assigned a score above 80 and different types of actions assigned a score below 80.

In addition, to compare with some of the currently popular approaches, the expertise of professional sports instructors labels the dataset. This enables supervised training and inference of scores using the proposed method. After labeling the dataset, the performance of different approaches on the smaller dataset, BGym, is evaluated. The evaluation metric used is Spearman correlation, commonly employed in popular approaches. As shown in Eq.equation 14. ρ is the Spearman's rank correlation coefficient. d_i is the difference between the ranks of corresponding values from the two samples. n is the number of samples.

$$Acc = \frac{\text{Successful matching actions}}{\text{Total number of actions}} \tag{12}$$

$$Rate80 = \frac{M + N}{\text{Number of action matches}} \tag{13}$$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{14}$$

4.2 DATASET

The AQA dataset (Parmar & Morris, 2019b) comprises sports videos from the Olympic Games, annotated with difficulty and scoring labels. However, the AQA dataset often suffers from frequent camera switches, resulting in inconsistent perspectives. Therefore, it is not suitable for addressing action matching problems in sports curriculum. To address this limitation, we propose the BGym dataset.

The BGym dataset is collected from the broadcast gymnastics course. Once this paper is accepted, the dataset will be made publicly available. The broadcast gymnastics routines were divided into eight sections: stretching exercises, chest expansion exercises, kicking exercises, lateral exercises, body rotation exercises, abdominal and back exercises, jumping exercises, and ending exercises. However, the final section is excluded, as some videos do not include it, leaving the template video with seven sections.

We selected one of the most standardized videos to serve as a template. The first row of Table 2 shows that the template video contains 7 actions, and the bottom row represents the 7 categories corresponding to the test video. Each exercise contains four eight-beat movements. We identified and collected the most representative movements from these eight-beat sequences, each lasting approximately 3 seconds. For instance, in the stretching exercise section, the selected three-second-long case includes flat-side stretching with hands, crouching with fists, hands-up stretching, and standing at attention. In addition, some beats in the four eight-beat movements are not executed with strict adherence to the standard. That is, certain beats exhibit significant deviations from the others. Consequently, three such instances are extracted from each practice session to serve as three videos in the dataset.

Table 2: Dataset overview

Number	Video category	Amount
-	Template Video	7
1	Lateral Exercises	24
2	Chest Expansion Exercises	21
3	Kicking Exercises	21
4	Stretching Exercises	24
5	Body Rotation Exercises	21
6	Abdominal and Back Exercises	21
7	Jumping Exercises	21

4.3 Parameter Ablation Experiment of MED module

We conducted ablation experiments to examine the benefits of different algorithms, as illustrated in Table 3. In these experiments, we maintained consistent conditions by utilizing seven template videos for template samples, 153 evaluation videos for test samples, and 3D features for the feature number. All schemes employed DTW with MED.

Initially, we evaluated the influence of the hyperparameter t in the MED module (as discussed in Section 3.2.1) on the experimental outcomes. The best overall performance was achieved when t was set to 0.1, resulting in Accuracy(Acc) and $Rate80$ values of 97.39% and 95.80%, respectively. The highest Acc value of 98.04% was obtained when t was 0.2. The MED module aids the algorithm in emphasizing the similarities and differences between pose features, thereby effectively evaluating motion quality.

Table 3: Ablation study of the parameter t

Method	t	Acc/%	Rate80/%
MED-DTW	0.1	97.39	**95.80**
	0.15	96.73	92.90
	0.2	**98.04**	65.00
	0.25	97.39	42.86

4.4 Ablation Experiment of the MED Module and Multi-feature Module

As shown in Table 4, the ablation experiment compares the performance of the MED module with the ED module. While the accuracy (Acc) remains consistent between the two methods, the $Rate80$ metric significantly improves with the MED module. This indicates the MED module's high discriminability and its ability to accurately differentiate variations in students's sports skills and abilities. The comparison results of the two methods are shown in Figure 4. We conduct a scoring display for the same character. In Figure 4(a), for videos with the same action, both methods yield the high scores that we expect. However, in Figure 4(b), with the template video fixed, although the score of ED-ACDTW drops slightly, it still does not meet the expectation of a low score as required in practical situations. MED-ACDTW assigns scores below 80 to all the non-corresponding videos, further demonstrating its discriminability.

When 3D features are used instead of 2D features, Acc decreases slightly by 0.7%, but $Rate80$ increases by approximately 11%. The integration of 2D and 3D features involves combining 2D centroid, angle, and orientation features with 3D orientation and upper and lower limb coordination features. This combination improves Acc by around 1% and enhances $Rate80$ by nearly 2% compared to using only 3D features.

The best performance is achieved by integrating all 2D and 3D features, including computing angle features in both dimensions. This configuration results in an Acc of 99.35% and a $Rate80$ of 95.33%. These results demonstrate that multi-feature fusion effectively captures diverse attributes

of moving characters from multiple perspectives, providing richer and more comprehensive information.

Table 4: Comparison of different features

Method	Feature	Acc/%	Rate80/%
ED-ACDTW	3D	96.73	16.99
MED-ACDTW	2D	97.39	82.17
	3D	96.73	93.27
	2D+ Partial 3D	98.69	95.24
	2D+3D	**99.35**	**95.33**

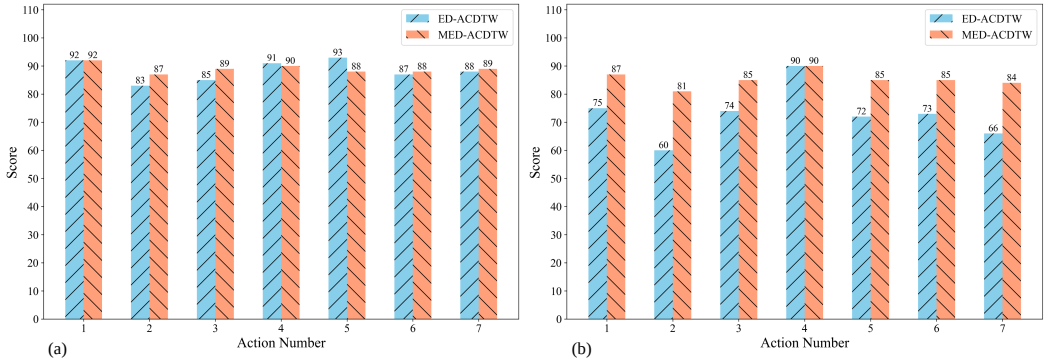

Figure 4: Comparing ED-ACDTW and MED-ACDTW methods. (a) shows the scores of the 7 test videos and the corresponding template video. (b) shows the scores of the 7 test videos and the stretching exercises template video.

4.5 Comparison of Different Methods

Table 5 presents the experimental findings, highlighting the challenges in assessing action quality within the dataset used in this study. To ensure a fair comparison, we re-implemented the evaluated methods based on the proposed MED framework. Since MED differs from the original distance modules, a "MED-" prefix is added to these re-implemented approaches. When all features are set to 3D, ACDTW demonstrates significant improvements compared to MED-DTW (Ying, 2021) and MED-Greedy (Effenberg et al., 2016) algorithms.

Table 5: Comparison of action methods

Method	Feature	Acc/%	Rate80/%
MED-DTW	3D	98.04	65.00
MED-Greedy	3D	94.12	79.93
MED-ACDTW	3D	96.73	93.27
	2D+3D	**99.35**	**95.33**

As shown in Table 5, although the Acc values are similar, the $Rate80$ for MED-ACDTW increases by approximately 28% over MED-DTW. This improvement indicates that MED-ACDTW's adaptive penalty mechanisms effectively address many-to-one and one-to-many mapping anomalies, refining the MED-DTW path. As a result, MED-ACDTW enables more distinct and reliable action quality assessments, meeting the requirements for usability in practical evaluation scenarios. We compare the effect of DTW paths and score result in the three algorithms in Figure 5.

In addition, this study compares MED-ACDTW with popular regression-based methods, as shown in Table 6. Two data-splitting strategies are employed. The first randomly selects 22 videos from the

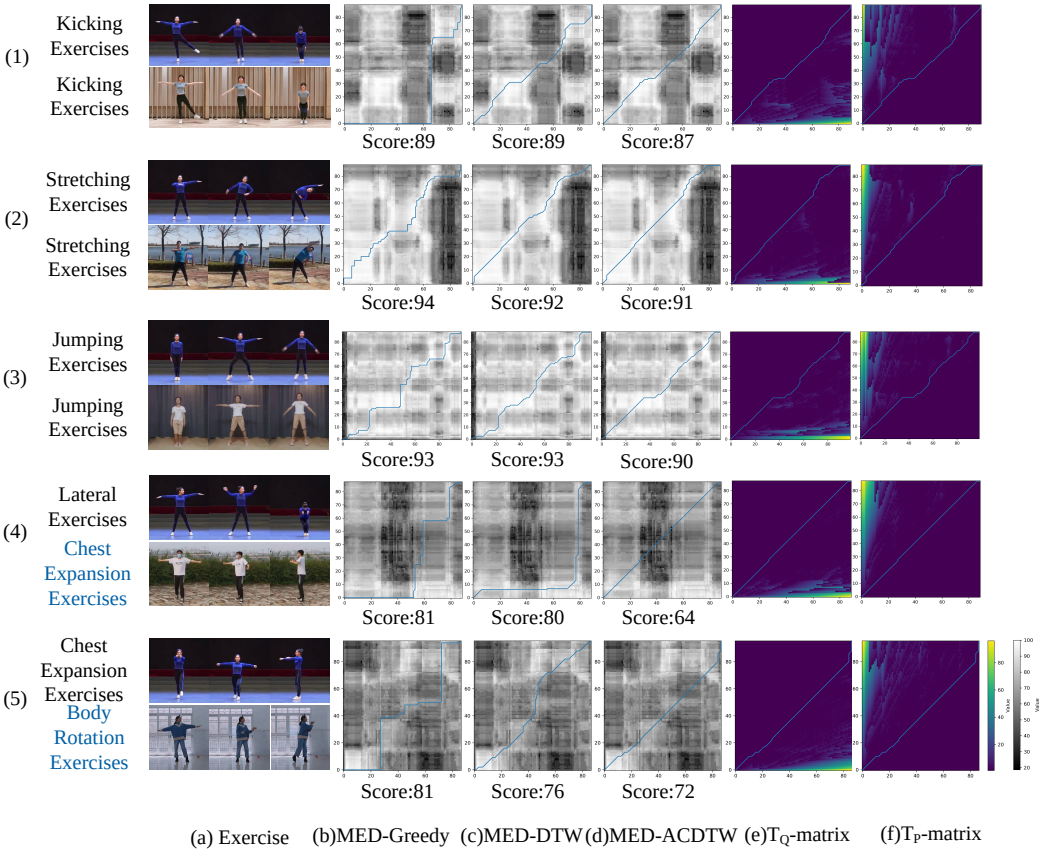

Figure 5: Action matching and score matrices. Columns (1), (2) and (3), the template videos and test videos are matched with the same action. In rows (4) and (5), the template videos and test videos are matched with different actions. Column (a) shows the individual's exercises. Columns (b), (c), (d), represent score paths or paths in different methods, namely the MED-Greedy, the MED-DTW, the MED-ACDTW. Columns (e) and (f) represent the penalty values of the T_Q matrix and the T_P matrix respectively..

Table 6: Comparison of regression methods

Method	Sp. Corr.	Sp. Corr.*
C3D-LSTM	0.3645	-
I3D-MLP (Tang et al., 2020)	0.9273	0.6699
I3D-DAE(Zhang et al., 2023)	0.9376	0.8465
HGCN(Zhou et al., 2023)	0.9064	0.8191
MED-ACDTW	**0.9418**	**0.9427**

dataset as the test set. The second strategy extracts three 3-second videos from a single action, using these as the test set, while ensuring the training set excludes corresponding actions for the specific character's movements. This results in a test set of 21 videos. In Table 6, the second strategy is denoted with an asterisk (*).

When the training set contains similar videos, MED-ACDTW achieves a Spearman correlation ($Sp.Corr.$) approximately 58% higher than the popular C3D-LSTM (Parmar & Tran Morris, 2017) method, 0.4% higher than I3D-DAE, and 4% higher than HGCN. Under the second splitting strategy, MED-ACDTW shows significant improvements, outperforming I3D-MLP by 28%, I3D-DAE by 10%, and HGCN by 12%. These results indicate that MED-ACDTW can achieve accurate scores even for untrained actions, provided template videos are available. In contrast, regression-based methods exhibit limited capability to perceive untrained actions. Figure 6 illustrates the performance of the methods in Table 6, showing their evaluated scores on the test set. Three specific points are selected for visualization. The curve of MED-ACDTW aligns more closely with the actual data, demonstrating better consistency and reliability.

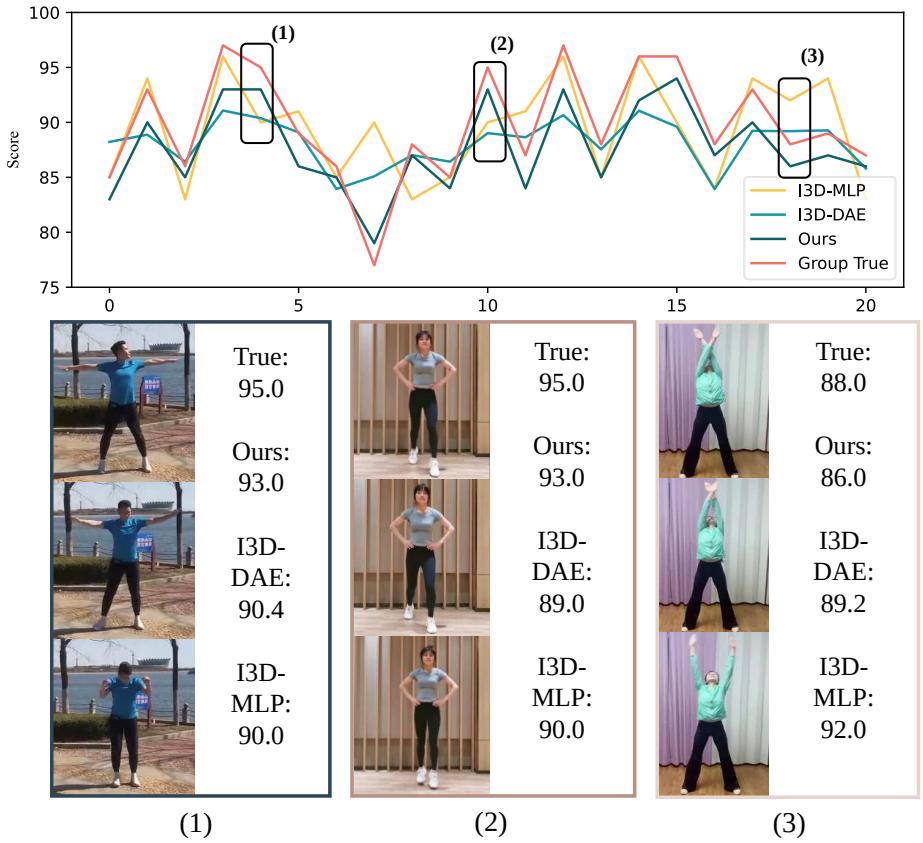

Figure 6: Comparison of popular methods. (1) represents lateral exercises. (2) represents jumping exercises. (3) represents abdominal and back exercises.

5 Conclusion

This study presents MED-ACDTW, a novel algorithm for action quality assessment. In terms of feature extraction, joint angles, barycentric features, directional features, and upper-lower limb linkage features are identified as key contributors to the scoring system. Additionally, the proposed training-free approach reduces dependency on dataset labeling, enhancing its adaptability to various scenarios.

The DTW algorithm is enhanced through modifications to the path score and the incorporation of adaptive constraints, significantly improving discriminability and mitigating one-to-many and many-to-one mapping issues in action time series. The MED-ACDTW framework effectively assesses action quality, demonstrating its potential for accurate and reliable evaluations. Future work will focus on refining existing features and introducing additional ones to better differentiate between similar movements. Moreover, we plan to enrich the dataset with more complex data, broadening the scope of the gymnastics dataset.ore complex data to expand our gymnastics dataset and extend our model to cover other sports categories.

ACKNOWLEDGMENTS

During the preparation of this work, we used LLMs in order to polish it. After using this tool, we reviewed and edited the content as needed and take full responsibility for the content of the publication.

REFERENCES

Khush Bakhat, Kashif Kifayat, M Shujah Islam, and M Mattah Islam. Katz centrality based approach to perform human action recognition by using omkz. *Signal, Image and Video Processing*, 17(4): 1677–1685, 2023.

Xiaobin Chang, Frederick Tung, and Greg Mori. Learning discriminative prototypes with dynamic time warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8395–8404, 2021.

Zexing Du, Di He, Xue Wang, and Qing Wang. Learning semantics-guided representations for scoring figure skating. *IEEE Transactions on Multimedia*, 2023.

Alfred O Effenberg, Ursula Fehse, Gerd Schmitz, Bjoern Krueger, and Heinz Mechling. Movement sonification: effects on motor learning beyond rhythmic adjustments. *Frontiers in neuroscience*, 10:219, 2016.

Shafkat Farabi, Hasibul Himel, Fakhruddin Gazzali, Md Bakhtiar Hasan, Md Hasanul Kabir, and Moshiur Farazi. Improving action quality assessment using weighted aggregation. In *Iberian Conference on Pattern Recognition and Image Analysis*, pp. 576–587. Springer, 2022.

Ye Gu, Ha Do, Yongsheng Ou, and Weihua Sheng. Human gesture recognition through a kinect sensor. In *2012 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1379–1384. IEEE, 2012.

Gökhan Güney, Talisa S. Jansen, Sebastian Dill, Jörg B. Schulz, Manuel Dafotakis, Christoph Hoog Antink, and Anne K. Braczynski. Video-Based Hand Movement Analysis of Parkinson Patients before and after Medication Using High-Frame-Rate Videos and MediaPipe. *Sensors*, 22 (20):7992, 2022. doi: 10.3390/s22207992.

Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5344–5352, 2015.

Huanhuan Li, Jingxian Liu, Zaili Yang, Ryan Wen Liu, Kefeng Wu, and Yuan Wan. Adaptively constrained dynamic time warping for time series classification and clustering. *Information Sciences*, 534:97–116, 2020.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, 2019.

Abrar Majeedi, Viswanatha Reddy Gajjala, Satya Sai Srinath Namburi GNVV, and Yin Li. Rica2: Rubric-informed, calibrated assessment of actions. In *European Conference on Computer Vision*, pp. 143–161. Springer, 2025.

Victoria Manousaki, Konstantinos Papoutsakis, and Antonis Argyros. Action prediction during human-object interaction based on dtw and early fusion of human and object representations. In Markus Vincze, Timothy Patten, Henrik I. Christensen, Lazaros Nalpantidis, and Ming Liu (eds.), *Computer Vision Systems*, pp. 169–179, Cham, 2021. Springer International Publishing.

Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1468–1476. IEEE, 2019a.

Paritosh Parmar and Brendan Tran Morris. 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*, pp. 304–313, 2019b.

Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 20–28, 2017.

Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pp. 556–571. Springer, 2014.

Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. Real time action recognition using histograms of depth gradients and random decision forests. In *IEEE winter conference on applications of computer vision*, pp. 626–633. IEEE, 2014.

Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9839–9848, 2020.

Yuan Wan, Xiao-Li Chen, and Ying Shi. Adaptive cost dynamic time warping distance in time series analysis for classification. *Journal of Computational and Applied Mathematics*, 319:514–520, 2017.

Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23967–23977, 2025.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2949–2958, 2022.

Xiang Ying. *Research on posture matching based on Openpose and its application in physical education*. PhD thesis, Nanjing Normal University, 2021.

Ling-An Zeng and Wei-Shi Zheng. Multimodal action quality assessment. *IEEE Transactions on Image Processing*, 2024.

Boyu Zhang, Jiayuan Chen, Yinfei Xu, Hui Zhang, Xu Yang, and Xin Geng. Auto-encoding score distribution regression for action quality assessment. *Neural Computing and Applications*, pp. 1–14, 2023.

Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. MediaPipe Hands: On-device Real-time Hand Tracking. https://arxiv.org/abs/2006.10214v1, 2020.

Shiyi Zhang, Sule Bai, Guangyi Chen, Lei Chen, Jiwen Lu, Junle Wang, and Yansong Tang. Narrative action evaluation with prompt-guided multimodal interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18430–18439, 2024.

Kanglei Zhou, Yue Ma, Hubert PH Shum, and Xiaohui Liang. Hierarchical graph convolutional networks for action quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.