ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Synthesize high-fidelity sports teaching demonstration videos using generative adversarial networks to enhance the level of action cognition

Jun Cao[1,*]

[1] School of Physical Education, Yuzhang Normal College, Nanchang, 330103, Jiangxi, China

SUMMARY: *To enhance the clarity and standardization of action demonstration resources in physical education teaching, this paper has constructed a high-fidelity sports teaching demonstration video synthesis model based on generative adversarial networks, and tested its impact on the level of action cognition from the perspective of teaching application. The experimental results show that the total score of action cognition in the experimental group increased from 72.1 points to 86.9 points, while that in the control group increased from 71.8 points to 78.6 points; the improvement rates of the experimental group in action sequence recognition and key point memory were 20.3% and 19.4% respectively, both higher than those of the control group. In the delayed test, the scores of the experimental group decreased from 86.9 points to 84.4 points, while those of the control group decreased from 78.6 points to 75.5 points. The retention effect was more stable, indicating that the high-fidelity synthesized demonstration videos can effectively promote students' understanding and memory of actions.*
Povzetek: This paper takes the insufficient quality of action demonstration resources in physical education classes as the starting point, and introduces a generative adversarial network to build a high-fidelity sports teaching demonstration video synthesis model. Through methods such as key point extraction, feature representation, and temporal constraints, the standardization and coherence of the demonstration videos are improved. Combined with teaching experiments, the synthesized videos were verified in terms of the improvement of action cognitive level and learning retention effect. The results show that such demonstration videos help students better grasp the structure and key points of the actions, and have promotional value in physical education practice.

KEYWORDS: *Generative Adversarial Network; Physical Education Teaching; Demonstration Video Synthesis*

1 Introduction

In physical education classes, whether students can understand the technical movements clearly largely depends on whether they can clearly see the movement structure, understand the sequence and remember the key steps during the demonstration. In actual teaching, teachers often need to demonstrate the same technical movement multiple times within a limited class period, while taking into account the different positions of students and the learning needs of students at different levels of foundation. Moreover, many ball games, gymnastics and rhythmic sports movements have the characteristics of strong continuity, frequent changes and numerous local details, which increases the cognitive load on learners when they only rely on live demonstrations [9]. Relying solely on on-site demonstrations is therefore prone to

*caojun7211@126.com
https://doi.org/10.65102/is2026540
Published: 30 April, 2026

problems such as restricted observation angles, insufficient exposure of details, and confusion in students' memory, making it even more difficult to correct mistakes in subsequent practice.

With the widespread adoption of video equipment and online platforms, recorded demonstrations, micro-videos, and online teaching segments have gradually been introduced into physical education classes to supplement on-site demonstrations and support students in repeatedly viewing and practicing with reference [6]. However, from the actual usage perspective, many teaching videos still have shortcomings in aspects such as shooting composition, light control, background interference, camera follow-up, and the standardization of demonstration actions, so they mainly serve the function of “classroom recording” and are difficult to become high-quality demonstration resources that can be repeatedly called upon in different grades and classes [9]. In addition, some earlier explorations of image- and video-based teaching in college physical education have even been retracted due to methodological or quality problems, which also reminds researchers to be cautious when constructing video resources for instruction [10]. For basic movements that require long-term training, the lack of a clear, stable, and highly standardized set of demonstration videos will further weaken the supporting role of video resources in movement cognition and skill transfer.

In this context, how to reconstruct and enhance the demonstration process based on existing action data, through means such as key point extraction, skeleton trajectory modeling and temporal constraints, and generate a high-fidelity sports teaching demonstration video that better meets the teaching requirements, has become a topic worthy of attention. Recent studies on generative adversarial networks show that such models can handle synthetic trajectories and spatiotemporal patterns with good flexibility, which provides a possible technical basis for reconstructing movement demonstrations [14, 20]. Conditional and task-oriented GAN variants have also been used to integrate domain knowledge or performance metrics into the generation process, further improving the controllability and utility of generated content [17, 19]. At the same time, there are also cases where GAN-based models have been retracted in other application domains, indicating that generation quality and validation procedures must be carefully examined before practical deployment [15]. Based on this, this paper, grounded in the practical needs of sports teaching for standardized, clear and repeatable demonstration resources, conducts research on the construction of a synthesis model for high-fidelity sports teaching demonstration videos, the optimization of temporal consistency and action authenticity, and the design of teaching experiments. It examines the teaching effect of the synthesized demonstration videos from aspects such as the total score of action cognition, changes in scores of different cognitive dimensions, and the retention over time, providing a feasible idea for the optimization and construction of demonstration resources in sports classrooms.

2 Literature Review

In recent years, research on sports teaching videos has mainly focused on the supporting role of video resources in action learning. Liu, for example, discussed the application of "micro-video + understanding teaching method" in college basketball courses, demonstrating that short, focused videos have strong adaptability in action explanation and classroom understanding [1]. Trabelsi et al. systematically reviewed the common practices of video modeling in physical education, pointing out that video demonstrations, repeated observation, and teaching feedback are important practical directions in current video-based teaching [2]. Ben Romdhane et al. compared the effects of different modalities and video control methods on basketball tactical learning, showing that the presentation of videos can affect learners' understanding of actions

and tactical processes [3]. Adams et al. discussed the value of video in sports teaching from the perspectives of online video application and video feedback, arguing that videos not only facilitate action observation but also contribute to subsequent skill learning and correction [4, 5]. These studies indicate that video resources have become an indispensable supporting means in sports teaching.

At the digital teaching level, Zou Yan and Han Yan discussed the application paths of short-video apps in college sports teaching, reflecting that short-video platforms are entering the sports classroom [7]. Wang Xiangyang's research on interactive object models, and Zhang Yangsheng et al.'s research on online video skill teaching systems, also indicate that sports teaching videos are evolving from single demonstration materials to systematic and interactive resources [8, 12]. Additionally, Cao Ling proposed that AIGC can empower vocational education teaching, although the research object is not the sports course, but its thinking shows that generative technologies have already had a realistic basis for entering the teaching scenario [11].

Compared with research on sports teaching videos, research related to generative adversarial networks is more focused on improving the quality of visual generation. Wibowo et al. found through comparative methods that GANs have good effects in enhancing the realism of character animations [16], indicating that such models have application potential in the generation of continuous visual content. Liu Jinming et al. combined knowledge bases with generative adversarial networks for multi-factor collaborative optimization [13]; Qi Han et al. proposed a hybrid quantum-classical WGAN for image generation [21]; Xu Zhenghua et al. improved GANs using attention contrastive learning for medical image modalities completion [18]. These studies show that generative adversarial networks have formed rich method accumulations in detail restoration, realism enhancement, and complex task adaptation.

However, existing research still has significant separations. One type of research focuses on how videos are used in sports teaching, paying attention to demonstration, feedback, and control methods; the other type of research focuses on how GANs generate more realistic visual content, focusing on model structure and generation effects. Studies that truly combine high-fidelity video generation with sports teaching and further examine its impact on action cognitive levels are still relatively rare. Therefore, how to use generative adversarial networks to synthesize high-fidelity sports teaching demonstration videos and verify their role in improving action cognition still has considerable research space.

3 Construction of a High-Fidelity Sports Teaching Demonstration Video Synthesis Model Based on Generative Adversarial Networks

3.1 Extraction and Representation of Sports Action Features

The generation effect of sports teaching demonstration videos is primarily determined by the accuracy of action feature extraction. Only by extracting the posture changes, joint relationships, and movement processes of the human body in the video can subsequent models better restore the standard actions and generate complete-structured and coherent demonstration videos. Therefore, before model training, the original sports action videos need to be characterized first, converting the image information into action sequences that can be used for learning.

Suppose a sports teaching video consists of T frame images, it can be represented as:

$$V = \{I_1, I_2, \dots, I_T\} \tag{1}$$

Among them, I_T represents the image of the t-th frame. Video is essentially the unfolding of continuous actions along the time dimension. Therefore, processing the video frame by frame is the basis for extracting action information.

Based on this, further, through the human keypoint detection method, the positions of the skeletal joints are extracted, and the human posture in each frame is represented as a set of key points. Thus, the entire video can be transformed into a sequence of skeletal postures:

$$S = \{P_1, P_2, \dots, P_T\} \tag{2}$$

Among them, P_T represents the human body posture information of the t-th frame. Compared with the original image, the skeleton sequence can more directly reflect the limb structure and movement changes, and can reduce the interference of external factors such as background and lighting on action recognition, making it more suitable for sports action analysis.

Only static posture information is not enough. Sports actions often have obvious continuity and rhythm, such as arm swinging, rotation, jumping, and landing processes, which all rely on the dynamic changes between consecutive frames to be reflected. Therefore, in the process of action representation, it is also necessary to combine the position changes between adjacent frames and the angle changes of some key joints to further describe the action process. This can not only reflect the action form at a certain moment, but also reflect the direction and amplitude of the action evolution.

To reduce the influence caused by different shooting conditions and individual differences, this paper uniformly processes the extracted key point data to make different samples be in a relatively consistent representation space. On this basis, the position features, motion change features, and joint structure features are fused to form an action comprehensive feature sequence:

$$F = \{F_1, F_2, \dots, F_T\} \tag{3}$$

Among them, F_T represents the comprehensive action feature of the t-th frame. This sequence retains both the spatial structure information of human postures and the temporal continuous information of action development, and can describe the standard demonstration actions in physical education teaching more completely.

In addition, the action demonstrations in physical education teaching usually include several key stages, such as the starting position, force application, transition, and ending. These stages are often the key points for learners to observe and understand the actions. Therefore, when representing the features, the key frames with more obvious changes in the actions should be highlighted, so that the model pays more attention to the core parts of the actions, thereby improving the authenticity and teaching relevance of the subsequent video generation.

3.2 Design of High-Fidelity Sports Teaching Demonstration Video Generation Model

After extracting and representing the characteristics of sports movements, a more stable and effective video generation model needs to be established on this basis. This model should not only be able to generate sports teaching demonstration videos with natural scenes and smooth movement transitions, but also should try to accurately present the basic structure, rhythm

changes and key technical links of the standard movements, so as to better meet the practical needs of teaching presentation and movement demonstrationl.

The model input mainly includes sports action video data, skeleton keypoint sequences, and action feature vectors extracted in the previous step. The original video data is used to provide the appearance information of the movement, the skeleton keypoint sequence is used to retain the human body posture structure, and the action feature vector further summarizes the rhythm changes and temporal relationships during the movement process. Through the joint input of multiple sources of features, the model can take into account both the movement form and the video performance effect during the generation process.

As shown in Figure 1, the high-fidelity sports teaching demonstration video generation model constructed in this paper mainly consists of the input data module, feature encoding module, generative adversarial learning module, and output module. The model first extracts key points and encodes action features from the sports action video to form the feature representation used for subsequent generation; then, the posture constraint signal and action rhythm signal are used as conditional inputs to guide the generation model to output video content that better meets the requirements of sports teaching; on this basis, the discrimination module evaluates the authenticity and action rationality of the generated results and returns the feedback results to the generation end to continuously improve the video quality.

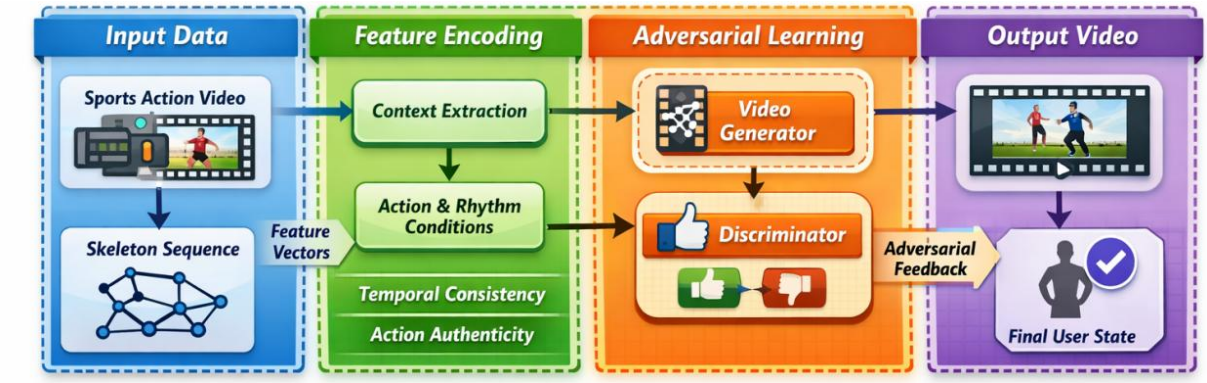

Figure 1: Flowchart of the high-fidelity sports teaching demonstration video generation model

During the generation stage, the model can be expressed as:

$$\hat{V} = G(F, z) \tag{4}$$

Among them, F represents the input sequence of action features, z represents the random noise vector, and V represents the generated high-fidelity sports teaching demonstration video. This expression indicates that the model output is not a mechanical replication of the original video, but reconstructs the video content under the constraint of action features, thereby ensuring that the generated result maintains the action norms while having good visual authenticity.

To ensure that the output video meets the requirements in both picture quality and action quality, this paper introduces a discriminator to conduct adversarial evaluation of the generation result. The discriminator not only judges whether the video is close to the real demonstration video, but also comprehensively identifies the naturalness of action connection and the rationality of local postures. Its adversarial objective can be expressed as:

$$\min_G \max_D V(D, G) = E_{V \sim p_{data}(V)}[\log D(V)] + E_{F \sim p(F), z \sim p(z)}[\log(1 - D(G(F, z)))] \tag{5}$$

In this process, the generator continuously improves the quality of the output video, and the discriminator continuously enhances the ability to distinguish between true and false. The two push the model to converge through repeated game.

3.3 Design of High-Fidelity Sports Teaching Demonstration Video Generation Model

In the generation of sports teaching demonstration videos, if only the clarity of a single frame is focused on, problems such as unstable connection between consecutive actions, obvious jumps in local postures, and distorted action rhythms are often encountered. For teaching demonstrations, these issues directly affect learners' observation and understanding of the action process. Therefore, during the training of the generation model, in addition to adversarial learning, it is necessary to further incorporate video temporal consistency constraints and action authenticity constraints, so that the output results not only have continuous and natural dynamic expressions but also can better restore the technical characteristics of standard sports actions.

The optimization of temporal consistency mainly involves reducing unreasonable abrupt changes between adjacent frames and maintaining smooth action transitions. Sports actions inherently have continuous evolving characteristics. Whether it is arm swinging, body rotation, or jumping and landing, there should be a relatively clear motion connection between the preceding and following frames. If the generated video shows obvious jitter or posture breaks at adjacent moments, it will weaken the readability of the demonstration video. Based on this consideration, this paper takes the differences between adjacent frames as the temporal constraint term, and its loss function can be expressed as:

$$L_{temp} = \sum_{t=2}^{T} \left\| \hat{I}_t - \hat{I}_{t-1} \right\|_2^2 \tag{6}$$

Among them, $\hat{I}_t$ represents the image result of the video at the t-th frame, and T represents the total number of frames in the video. This formula reflects the variation amplitude between adjacent frames. The smaller the loss value, the more stable the transition between the previous and subsequent frames, and the better the video continuity.

Using only the temporal constraints at the pixel level is not sufficient, because the demonstration of sports actions also pays more attention to whether the change process of human body posture is reasonable. Therefore, this paper further conducts temporal control at the skeleton level and constrains the movement trend of key points in adjacent frames to remain consistent. Let P^t be the representation of the human body posture key points corresponding to the t-th frame of the generated video, then the skeleton temporal consistency loss can be written as:

$$L_{pose} = \sum_{t=2}^{T} \left\| \hat{P}_t - \hat{P}_{t-1} \right\|_2^2 \tag{7}$$

This method can more directly restrict the trajectory of human body movements, avoiding situations where the video may have smooth visuals but the actual posture changes deviate from the logic of sports movements. For sports teaching demonstration videos, this constraint is particularly important because students focus on the details of joint extension, body transfer, and force application paths when watching the demonstration.

The optimization of action authenticity is mainly used to improve the consistency between

the generated results and the standard movements. Sports teaching demonstration videos are different from general entertainment videos. Their value is not only "like real", but also lies in "the movements are correct". Therefore, this paper incorporates the feature differences between the generated actions and the real demonstration actions into the loss function to control the key technical actions from deviating. Let the sequence of features of the real demonstration action be F, and the sequence of action features extracted from the generated video be $\hat{F}$, then the action authenticity constraint can be expressed as:

$$\mathrm{L_{act}} = \left\|\mathrm{F} - \hat{\mathrm{F}}\right\|_2^2 \tag{8}$$

This formula compares the distances of the real actions and the generated actions in the feature space, enabling the model to retain the structural features, technical rhythm, and amplitude changes of the standard actions as much as possible during the training process, thereby improving the teaching applicability of the demonstration videos.

When generating teaching demonstration videos in practice, the temporal coherence and the authenticity of the actions are not mutually exclusive. The former addresses whether the actions are smoothly connected from one to another, while the latter focuses on whether the action postures,Is the force application method and technical structure accurate. Only when both aspects are ensured can the generated video truly possess watchability and teaching value. If one solely pursues smooth transitions in the visuals, the resulting outcome often tends to be overly smooth, losing the necessary force and distinctiveness of the actions; if only the action reproduction is emphasized, it may cause awkward frame connections, affecting the overall viewing effect. Based on this consideration, in the model training process of this paper, no single indicator was given special emphasis. Instead, the adversarial loss, temporal consistency loss, and action authenticity loss were incorporated into a unified optimization process to balance the coherent expression of the video and the accurate presentation of the action demonstration. The total loss function is expressed as:

$$\mathrm{L} = \lambda_1 \mathrm{L_{adv}} + \lambda_2 \mathrm{L_{temp}} + \lambda_3 \mathrm{L_{pose}} + \lambda_4 \mathrm{L_{act}} \tag{9}$$

Among them, The adversarial loss is defined by L_{adv}, the temporal consistency loss is L_{temp}, the temporal constraint loss is L_{pose}, and the action authenticity loss is named L_{act}. Additionally,we have introduced four parameters: $\lambda_1, \lambda_2, \lambda_3,$ and λ_4,which correspond to the weight factors of each loss component.These parameters enable us to appropriately sacrifice and coordinate the relationship between action coherence and action normativity while maintaining the visual quality of the video.

From the training results, it can be seen that after introducing these improvement factors, the generated videos will have improvements in the transformation between adjacent frames, the smoothness of limb movements, and the degree of reverting important actions.

Especially in the context of sports teaching, the motion capture results of students often require continuous visual effects and clear body movement trajectories to determine whether the movements are correct. Improving the time synchronization and realism can reduce the possibility of unnecessary images and redundant interfering factors, enabling the generated images to meet the requirements for teaching use.

3.4 Generation Process of Synthetic Demonstration Videos

After all the motion behavior features have been extracted and the corresponding models have been constructed, the stage of adjusting the timing and the authenticity of the actions to obtain a

better synthesized demonstration video begins. "Data acquisition - feature representation - constraint conditions - video generation - evaluation feedback - iterative correction - demonstration video" is the core of this stage. Its purpose is to make the generated film achieve the best effect in terms of motion construction, shot coherence, and teaching adaptability.

From the perspective of steps, the first step is to read the original sports video and the corresponding joint array, and uniformly organize them into action samples; the second step is to extract the motion feature codes from the action samples, including the coordinates of each joint point and motion states and motion speeds and other motion features. Let the input action feature sequence be f, and the random noise vector be z, then the synthesized video generated by the generator can be expressed as:

$$\widehat{\mathrm{V}} = \mathrm{G}(\mathrm{F}, \mathrm{z}) \tag{10}$$

Among them, $\hat{V}$ is a sports teaching demonstration video generated by the model. This model indicates that the generated video is not simply a random mixture of noise, but is constrained by the characteristics of the movement. Based on this, the desired target video is constructed, so its output can better represent the standard process of sports movement techniques.

The video generated by the generator is input to the discriminator to determine the authenticity and quality of the video, and to determine the action continuity and whether the key postures are correct. At the same time, the evaluation results can be saved. As for:

$$\mathrm{D}(\widehat{\mathrm{V}}) \in [0,1] \tag{11}$$

Among them, $D(\hat{V})$ is used to represent the judgment result of the discriminator regarding the authenticity of the generated video. The value is closer to 1, indicating that the generated video is more similar to the real teaching demonstration video in terms of picture presentation, action completion degree, and overall presentation. As the training progresses, the feedback information output by the discriminator will continuously be sent back to the generator, enabling it to adjust the action details, frame-to-frame connection, and video structure, thereby gradually improving the stability and realism of the generated results.

In order to ensure that the final generated video not only "looks like" but also "acts accurately and has a coherent process", during the training process, it is not sufficient to only rely on a single objective for constraints. Instead, the adversarial loss, temporal consistency loss, and action authenticity loss should all be taken into consideration. Based on this, this paper jointly introduces these three types of losses into the generation stage and constructs a joint optimization objective. Its expression is as follows:

$$\mathrm{L} = \lambda_1 \mathrm{L}_{\mathrm{adv}} + \lambda_2 \mathrm{L}_{\mathrm{temp}} + \lambda_3 \mathrm{L}_{\mathrm{act}} \tag{12}$$

Among them, L_{adv} is the adversarial loss, L_{temp} is the temporal consistency loss, L_{act} is the action authenticity loss, and λ_1, λ_2, λ_3 are the weight coefficients of each term. During model training, by minimizing the total loss function, the generated video gradually approaches the performance form of the real physical education teaching demonstration video.

Specifically at the implementation level, the generation of the demonstration video can be further broken down into the following algorithmic steps.

Algorithm 1: Steps for generating synthetic demonstration videos

Input: Sports action video sample V, skeleton keypoint sequence S, action feature sequence F, random noise z

Output: High-fidelity sports teaching demonstration video $\widehat{\mathrm{V}}$

Steps 1: Data input and preprocessing

Read the sports action video sample, extract video frames and keypoint sequences, normalize, crop, and uniformly adjust the duration of the sample to form standardized input data for model training.

Step 2: Action feature encoding

Based on the skeleton keypoint sequence and video frame information, extract action position features, rhythm features, and structural features, and construct a comprehensive action feature sequence F.

Step 3: Conditional constraint injection

Inject posture constraint information, action category information, and rhythm information as conditional inputs, and send them together with the action feature sequence to the generator to enhance the model's control ability over standard sports actions.

Step 4: Generate demonstration video

Use the generator to calculate the mapping relationship from action features to the video space, and output the initial synthetic demonstration video:$\widehat{\mathrm{V}}^{(\mathrm{k})} = \mathrm{G}(\mathrm{F}, \mathrm{z})^{(\mathrm{k})}$

Among them, k represents the result generated in the k-th iteration.

Step 5: Authenticity Discrimination

The generated video $\widehat{\mathrm{V}}^{(\mathrm{k})}$ and the real demonstration video are jointly input into the discriminator to obtain the authenticity score and feedback information, which is used to evaluate whether the current result meets the expected standard.

Step 6: Loss Calculation and Parameter Update

The total loss value is calculated based on the adversarial loss, temporal consistency loss, and action authenticity loss, and the parameters of the generator and discriminator are updated using the backpropagation method. The parameter update process can be expressed as:$\theta_{\mathrm{k}+1} = \theta_{\mathrm{k}} - \eta\nabla\mathrm{L}(\theta_{\mathrm{k}})$

Among them, θ_k represents the model parameters at the k-th iteration, η represents the learning rate, and $\nabla L(\theta_k)$ represents the gradient of the loss function with respect to the parameters.

Step 7: Iterative Optimization and Result Output

The process from Step 4 to Step 6 is repeated until the model loss converges or reaches the preset number of training iterations. Finally, the high-fidelity sports teaching demonstration video V^ is output.

To provide a more intuitive illustration of the overall flow of the demonstration video generation process, it can be summarized as follows: $(\mathrm{V}, \mathrm{S}) \rightarrow \mathrm{F} \rightarrow \mathrm{G}(\mathrm{F}, \mathrm{z}) \rightarrow \widehat{\mathrm{V}} \rightarrow \mathrm{D}(\widehat{\mathrm{V}}) \rightarrow \mathrm{Optimize}$

This process indicates that the model takes the original video and the skeleton sequence as inputs, first extracting and encoding the action information to obtain the corresponding feature representation, then the generator combines the features with random variables to generate the demonstration video, subsequently sending the generated result to the discriminator for authenticity judgment, and based on the feedback results, continuously adjusting the model parameters. Through continuous iterations, a high-quality and accurately expressed teaching demonstration video is ultimately obtained.

3.5 Model Training Strategy and Stability Analysis

The high-fidelity sports teaching demonstration video synthesis model is trained using a generative adversarial structure. The core idea is to continuously "identify errors" by the discriminator, pushing the generator to output video sequences that are closer to the real demonstrations. Let G_θ represent the generator and D_θ represent the discriminator. The real demonstration video segments are denoted as $\mathrm{x} \sim \mathrm{p}_{\mathrm{real}}$, and the synthesized video segments are

$\tilde{x} \sim= G_\theta G(z, s)$, where z is a random noise vector and s is the conditional feature extracted from the sequence of action key points. The adversarial loss is in the form of cross-entropy, and the optimization objective of the discriminator is:

$$\mathcal{L}_{\mathrm{D}} = -\mathbb{E}_{\mathrm{x}\sim \mathrm{p}_{\mathrm{real}}}[\log \mathrm{D}_{\theta_{\mathrm{D}}}(\mathrm{x}) - \mathbb{E}_{\tilde{\mathrm{x}}\sim \mathrm{p}_{\mathrm{G}}}[\log(1 - \mathrm{D}_{\theta_{\mathrm{D}}}(\tilde{\mathrm{x}}))\mathrm{g}] \tag{14}$$

The generator's objective on the adversarial term is:

$$\mathcal{L}_{\mathrm{adv}}^{\mathrm{G}} = -\mathbb{E}_{\tilde{\mathrm{x}}\sim \mathrm{p}_{\mathrm{G}}}[\log \mathrm{D}_{\theta_{\mathrm{D}}}(\tilde{\mathrm{x}})] \tag{15}$$

Based on this, the temporal consistency loss $\mathcal{L}_{\mathrm{tc}}$ and the pose authenticity (or key point re-projection) loss $\mathcal{L}_{\mathrm{pose}}$ defined earlier are also considered. The total loss of the generator is written as:

$$\mathcal{L}_{\mathrm{G}} = \lambda_{\mathrm{adv}} \mathcal{L}_{\mathrm{adv}}^{\mathrm{G}} + \lambda_{\mathrm{tc}} \mathcal{L}_{\mathrm{tc}} + \lambda_{\mathrm{pose}} \mathcal{L}_{\mathrm{pose}} \tag{16}$$

Among them, λ_{adv}, λ_{tc}, λ_{pose} are weight coefficients used to balance the realism of the scene, the smoothness of the time sequence, and the accuracy of the action skeleton. During training, small-batch stochastic gradient descent is alternately updated the parameters of the discriminator and the generator. The parameter update form can be written as:

$$\theta_{\mathrm{D}} \leftarrow \theta_{\mathrm{D}} - \eta_{\mathrm{D}} \nabla_{\theta_{\mathrm{D}}} \mathcal{L}_{\mathrm{D}} \tag{17}$$

Among them, η_D, η_G are the learning rates of the discriminator and the generator, respectively.

To make the training process more stable and avoid common phenomena such as "oscillation" and "mode collapse" in adversarial training, several processing steps have been made in the training strategy. (1) The discriminator uses spectral normalization or weight constraints to suppress large gradients and reduce training oscillations; (2) A slight smoothing is applied to the real sample labels to weaken the discriminator's overconfidence in suppressing the generator; (3) The weights of $\mathcal{L}_{\mathrm{tc}}$ and $\mathcal{L}_{\mathrm{pose}}$ are adjusted in stages: in the first few training rounds, the adversarial loss is more emphasized, ensuring the overall scene formation, and then gradually increase the weights of temporal and pose constraints to make the model focus on the coherence of the action and the details of the skeleton. In actual training, the batch size is kept at a small level to update parameters more frequently, and the update ratio of the generator and the discriminator is 1:1 to prevent one side from training too quickly and making the other side difficult to keep up.

Based on the above design, the training process of this section can be summarized as the following algorithm steps.

Algorithm 2 Training Flow of the High-Fidelity Sports Teaching Demonstration Video Generation Model

Input: Sports teaching demonstration video dataset $\{x_i\}$, corresponding action key point sequence $\{k_i\}$, initialized parameters θ_G, θ_D, learning rates η_G, η_D, loss weights λ_{adv}, λ_{tc}, $\mathcal{L}_{\mathrm{pose}}$, number of training epochs MaxEpoch. Output: Generator parameters θ_{G^*} after training convergence.

1. Randomly initialize the parameters of the generator and discriminator θ_G, θ_D, set the learning rates and loss weights.

2. In each training round, randomly sample several segments of real demonstration videos

$\{x_i\}$ from the dataset, and extract the corresponding key point sequences $\{\mathrm{x_i}\}$, construct the conditional feature s_i.

3. Sample noise vector z_i for each sample, and use the generator to obtain the synthesized demonstration segment $\tilde{x}_i = G_\theta G(z_i, s_i)$.

4. Fix the generator parameters, calculate the discriminator loss L_D using the real samples $\{x_i\}$ and the synthesized samples $\{\tilde{\mathrm{x}}_\mathrm{i}\}$, update the discriminator parameters according to $\theta_D \leftarrow \theta_D - \eta_D \nabla_{\theta_D} \mathcal{L}_D$, and consider adding spectral normalization or gradient clipping as necessary. Fix the discriminator parameters, regenerate the synthetic fragment $\tilde{\mathrm{x}}_\mathrm{i}$, calculate $\mathcal{L}_{\mathrm{adv}}^{\mathrm{G}}$, $\mathcal{L}_{\mathrm{tc}}$, and obtain the total generator loss L_G.

5. Update the generator parameters according to $\theta_\mathrm{D} \leftarrow \theta_\mathrm{D} - \eta_\mathrm{D} \nabla_{\theta_\mathrm{D}} \mathcal{L}_\mathrm{D}$, and perform necessary clipping on the gradients to prevent gradient explosion.

7. Adjust the relative sizes of λ_{adv}, λ_{tc}, λ_{pose} based on the training progress. In the early stage, focus on overall adversarial learning, and gradually strengthen temporal and pose constraints in the later stage.

8. Periodically evaluate metrics such as structural similarity, temporal continuity, and key point error on the validation set. If there is no significant improvement for a long time, appropriately reduce the learning rate or terminate the training prematurely.

9. When reaching the preset number of iterations or meeting the convergence criterion, output the final parameters θ_{G^*} of the generator.

The aforementioned training method not only enables the model to complete the adversarial training process more stably, but also takes into account aspects such as realism, continuity, and skeleton accuracy in the design of the loss term and weights. This provides the possibility for later application of the generated demonstration videos in the teaching environment.

3.6 Quality Evaluation Indicators and Feedback Mechanism for Synthetic Demonstration Videos

In terms of objective evaluation, the peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are first used to measure the degree of similarity between the synthesized video and the real demonstration video at the picture level. Let the real video frame be x, the synthesized video frame be $\tilde{\mathrm{x}}$, and the frame size be $H \times W_H$. Then, the mean square error is:

$$\mathrm{MSE} = \frac{1}{\mathrm{HW}} \sum_{\mathrm{i}=1}^{\mathrm{H}} \sum_{\mathrm{j}=1}^{\mathrm{W}} \left(\mathrm{x_{ij}} - \tilde{\mathrm{x}}_{\mathrm{ij}}\right)^2 \tag{18}$$

The peak signal-to-noise ratio is defined as:

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}} \right) \tag{19}$$

Among them, MAX represents the maximum pixel value (such as 255). The structural similarity is in its classic form:

$$\mathrm{SSIM}(\mathrm{x}, \tilde{\mathrm{x}}) = \frac{(2\mu_\mathrm{x}\mu_{\tilde{\mathrm{x}}} + \mathrm{C}_1)(2\sigma_{\mathrm{x}\tilde{\mathrm{x}}} + \mathrm{C}_2)}{(\mu_\mathrm{x}^2 + \mu_{\tilde{\mathrm{x}}}^2 + \mathrm{C}_1)(\sigma_\mathrm{x}^2 + \sigma_{\tilde{\mathrm{x}}}^2 + \mathrm{C}_2)} \tag{20}$$

The average is calculated through a sliding window over the time sequence to obtain the structural similarity score of the entire demonstration video, which is used to reflect the fidelity

of the picture details and local structure.

Considering that the sports action itself has obvious skeletal structure features, relying solely on picture similarity is difficult to reflect whether the action is "performed correctly" or "positioned correctly". Therefore, the position error of key points is introduced as an indicator of action reconstruction accuracy. Let the true key point sequence be {pt(n)}, and the corresponding key point sequence of the synthesized video be {p~t(n)}, where t is the time step, n is the joint number, and N is the total number of joints. Then, the average key point error is denoted as:

$$\mathrm{E_{kp}} = \frac{1}{\mathrm{TN}} \sum_{\mathrm{t=1}}^{\mathrm{T}} \sum_{\mathrm{n=1}}^{\mathrm{N}} \left\| \mathrm{p_t^{(n)}} - \tilde{\mathrm{p}}_\mathrm{t}^{(\mathrm{n})} \right\|_2 \tag{21}$$

The smaller this indicator is, the smaller the deviation of the generated video from the joint positions and motion trajectories will be, and it will be closer to the real demonstration.

At the temporal level, to depict the continuity of the action, the inter-frame change difference is introduced as an auxiliary indicator. Let the difference between adjacent frames in the real video be $\Delta x_t = x_t - x_{t-1}$, and the difference between adjacent frames in the synthesized video be $\Delta \tilde{\mathrm{x}}_\mathrm{t} = \tilde{\mathrm{x}}_\mathrm{t} - \tilde{\mathrm{x}}_{\mathrm{t-1}}$. Then, the temporal stability loss can be written as:

$$\mathcal{L}_{\mathrm{temp}} = \frac{1}{\mathrm{T}-1} \sum_{\mathrm{t=2}}^{\mathrm{T}} \left\| \Delta \mathrm{x_t} - \Delta \tilde{\mathrm{x}}_\mathrm{t} \right\|_1 \tag{22}$$

The smaller this indicator is, the lower the error in joint position reconstruction and action trajectory representation of the generated video will be, and the closer it is to the actual demonstration action.

Based on the above objective indicators, a normalized quality score Q_{obj} is constructed comprehensively, for example:

$$\mathrm{Q_{obj}} = \mathrm{w_1}\ \widetilde{\mathrm{SSIM}} + \mathrm{w_2}\ \widetilde{\mathrm{PSNR}} + \mathrm{w_3}\ (1 - \widetilde{\mathrm{E_{kp}}}) \tag{23}$$

Among them, ˜ represents the normalized indicator values, and $\mathrm{w_1}$, $\mathrm{w_2}$, $\mathrm{w_3}$ are the weight coefficients used to balance the contributions of picture details, overall clarity, and accuracy of the action skeleton.

We invited some experienced physical education teachers and a group of students to participate in the subjective evaluation stage. They were asked to watch our overall introduction video and give corresponding scores, including "clarity", "standardness of movements" and "understandability of teaching", with ratings ranging from 1 to 5, from lowest to highest. Teachers paid more attention to whether the technical movements conformed to the teaching materials and assessment standards, while students were more concerned about whether they could "understand, remember, and follow", and together they formed the subjective quality score Q_{sub}. Through a simple linear combination:

$$\mathrm{Q_{total}} = \alpha\, \mathrm{Q_{obj}} + \beta\, \mathrm{Q_{sub}} \tag{24}$$

The comprehensive quality index for practical teaching applications can be obtained, where α and β are used to adjust the weights between the objective indicators and subjective feelings.

During the training and parameter adjustment stage, these indicators are not only used for the final effect evaluation, but also for forming a closed-loop feedback mechanism: when the

objective indicators show "clear picture but large errors in key points", the weight of the pose constraint term should be appropriately increased; when the teacher feedbacks "the movements are continuous but the rhythm is unnatural", the focus should be on checking the loss of temporal consistency and the temporal structure of the generator; when the students report "can see clearly but not easy to remember", adjustments need to be made in the selection of generated samples and the way of action decomposition. After continuous iterations involving indicator assessment, result feedback, and parameter adjustment, the generated sports teaching demonstration videos can achieve a more harmonious effect in terms of picture presentation, action reproduction, and teaching application.

4 Data Collection and Model Parameters

4.1 Source of Sports Action Video Data

To ensure the stability of subsequent model training and the usability of the generated results, this paper mainly considers the standardization of actions, video clarity, diversity of sample types, and teaching adaptability when selecting data. The sports action videos used mainly come from standardized teaching demonstration resources, public action video resources, and supplementary collected samples during the research process. This processing approach facilitates the model's learning of more standardized action structures and enhances its adaptability to different action forms and shooting conditions.

Standardized teaching videos are mainly used to provide clear, complete rhythm, and highly demonstrative samples. These videos usually can present the starting, exertion, completion, and ending processes of the action in a complete manner, making them an important basis for the model to learn standard actions. Public action video resources are used to expand the sample size and enrich the types of actions, allowing the model to encounter more different angles, different backgrounds, and different movement states during training. In addition, based on the research content of this paper, some typical sports teaching actions were also supplemented and collected to improve the degree of alignment between the data and the specific teaching scenarios.

During the sample selection process, this paper mainly follows the following principles: First, the action process is complete, and it can clearly present the changes in the action; second, the video picture quality is good, and it tries to avoid the influence of blurriness, occlusion, and shaking on the extraction of key points; third, the action category has certain representativeness, covering the common basic actions in sports teaching; fourth, different perspectives and different expression forms of video samples are retained as much as possible to enhance the richness of the data foundation.

According to the above sources and selection principles, the data sources of sports action videos are organized as follows.

Table 1: Data Sources of Sports Action Videos

Data source category	Data content	Main action types	Sample characteristics	Main purpose
Standardized teaching videos	School physical education demonstration videos; standardized movement demo videos	Radio calisthenics, basketball shooting, volleyball serving, arm swinging in running, etc.	Movements are relatively standardized, rhythm is clear, with strong demonstrative value	Used to learn standard movement structures and demonstration procedures
Public action video resources	Public video samples; action-recognition-related video clips	Jumping, arm swinging, body rotation, squatting, kicking, etc.	Large number of samples, diverse action types, varied shooting conditions	Used to expand training samples and enhance model adaptability
Supplementary collected samples	Action videos collected by the researcher during the study	Basketball shooting, volleyball serving, basic gymnastics movements, etc.	Highly aligned with the research content; viewpoint and background easy to control	Used to supplement key action categories and improve task specificity
Skeletal keypoint data	Skeletal sequences and pose information extracted from raw videos	Sequences of keypoint changes corresponding to various actions	Better preserves movement structure and is less affected by background interference	Used for motion feature extraction and as model input

As can be seen from Table 1, the data used in this paper is not from a single source. Instead, it combines normative demonstration videos, public video resources, and supplementary collected samples for use. This not only ensures that the model learns relatively standard sports movements, but also avoids the limitations caused by overly single sample sources. At the same time, the skeleton key point data is further extracted from the original videos and plays a fundamental role in the subsequent feature encoding and model training.

4.2 Model Parameter Settings

To ensure the stability of the model training process, this paper, in accordance with the characteristics of the video generation task, has uniformly set the input specifications, training parameters, and loss weights. The input video resolution is set to 256×256, the length of a single video segment is set to 32 frames, the number of skeleton key points is 17, and the dimension of the action feature is set to 128. In the training stage, the Adam optimizer is used, the initial learning rate is set to 0.0002, the batch size is set to 8, and the number of training rounds is set to 200. During the training phase, the generator and the discriminator alternate iterations at a ratio of 1:1 to prevent one side from updating too quickly and affecting the overall training stability.

Table 2 summarizes the model parameter settings, mainly including input parameters,

training parameters, network parameters, and loss parameters. According to the experimental goals, the adversarial loss weight is set to 1.0, the temporal consistency loss weight is set to 0.6, and the action authenticity loss weight is set to 0.8. This parameter configuration is mainly to achieve a reasonable balance among video realism, smoothness of action transitions, and accuracy of technical action expression.

Table 2: Model Parameter Settings

Parameter category	Parameter name	Parameter value
Input parameters	Input video resolution	256×256
Input parameters	Frames per video clip	32 frames
Input parameters	Number of skeletal keypoints	17
Input parameters	Dimension of motion features	128-D
Training parameters	Batch size	8
Training parameters	Number of training epochs	200
Training parameters	Initial learning rate	0.0002
Training parameters	Optimizer	Adam
Training parameters	Adam parameter β1	0.5
Training parameters	Adam parameter β2	0.999
Network parameters	Generator update steps	1 per epoch
Network parameters	Discriminator update steps	1 per epoch
Loss parameters	Weight of adversarial loss	1.0
Loss parameters	Weight of temporal consistency loss	0.6
Loss parameters	Weight of motion authenticity loss	0.8

As can be seen from Table 2, the main conditions set in this experiment took into account how to maintain the stability and consistency of action execution and duration to meet the requirements of video recording. Such a setup can ensure the generation of high-quality teaching demonstration videos and provide a basic model for further analyzing the quality of synthesized videos and studying the improvement of action understanding effects.

5 Evaluation of the Teaching Application Effect of High-Fidelity Sports Teaching Demonstration Videos

5.1 Evaluation of the Quality of Synthetic Videos

In order to assess the quality of the synthesized teaching videos, this experiment comprehensively evaluated the quality of the synthesized videos from five dimensions (similarity, resolution, length, reproduction degree, and teachability): in addition to comparing the similarity between the synthesized video and the original demonstration video, its applicability in practical movements was also considered. Since a single dimension cannot fully reflect the quality of a work, the combined comparison method, error analysis method, diachronic analysis method, and verb type comparison method were adopted to evaluate the quality of the combined videos.

From the overall indicators, the synthetic video shows a good degree of proximity in all quality indicators. The scores for structural similarity and clarity are relatively high, indicating that the generated video is already quite close to the real video in terms of picture organization, action outline, and local details; the temporal continuity and posture restoration degree are slightly lower than those of the real video, but the overall gap is not large, indicating that the

model can maintain the continuity of the action process and the basic posture structure well. Especially in the demonstrations of actions with fixed rhythms, the synthesized videos have been able to accurately reflect the main technical aspects. As shown in Figure 2, the four key indicators indicate that the morphological characteristics of the real video and the synthesized video are basically the same. The configuration similarity and image clarity of the synthesized video are closer to those of the real video, but the imitation accuracy of the synthesized video is still not ideal. This also indicates that this model has a relatively stable ability to maintain the overall movement framework, but for certain minor joint details, further improvement is still needed.

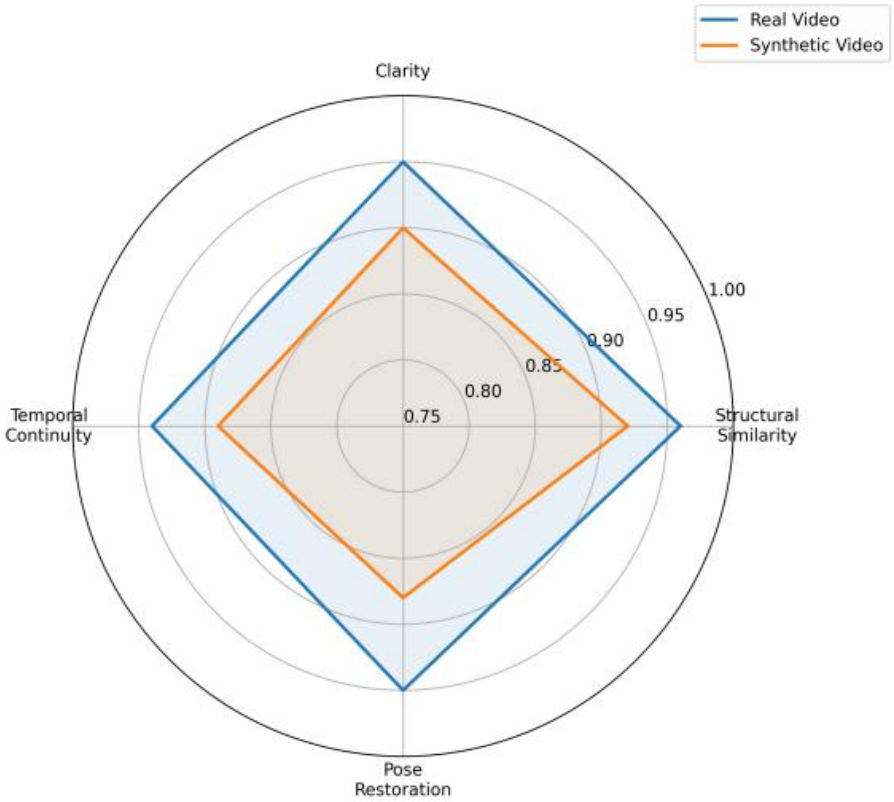

Figure 2: Radar Chart of Quality Indicators for Synthetic Video and Real Video

In terms of local motion error, there are certain differences in the reconstruction effects at different joint positions. Generally speaking, the errors at relatively stable parts such as the head, shoulders, and hips are smaller, while the errors at parts with larger motion amplitudes and faster change speeds such as the elbows, wrists, knees, and ankles are more obvious. This result is in line with the actual characteristics of sports movements, because the extremities of the limbs change most frequently during swinging, exertion, and turning, and require higher temporal modeling ability of the generation model. As shown in Figure 3, the deviation distribution of key points at the head and shoulders is relatively concentrated, with a lower median, indicating that the model has relatively stable control over the proximal parts of the upper limbs and the main trunk of the body; while the box range of the wrists, knees, and ankles is larger, and the dispersion is higher, indicating that these parts are more prone to fluctuations during the generation process. Especially the ankle, its deviation value and dispersion are relatively higher, reflecting that rapid lower limb movements are still a difficult part to handle in the current model.

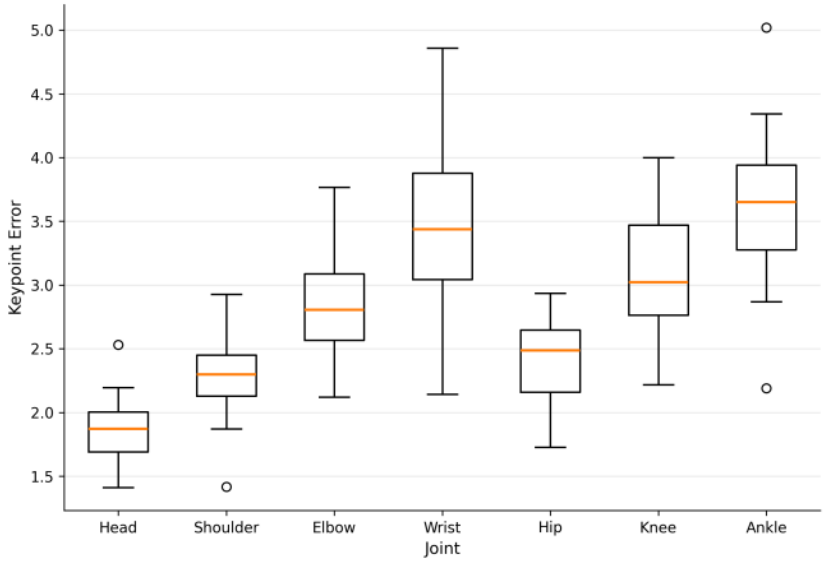

Figure 3: Box plot of deviation distribution of key points at different joints

From the temporal performance perspective, the change trend of the synthesized video throughout the entire action process is basically consistent with that of the real video. Whether in the starting stage, the action execution stage, or the ending stage, the inter-frame difference curve of the generated video can well follow the changing rhythm of the real video. This indicates that the model has achieved a certain stability in the time dimension and can reasonably organize the action process without large-scale frame skipping or action breaks. As shown in Figure 4, the overall trend of the curves of the real video and the synthesized video is close, both showing fluctuations and rises in the middle action execution stage, and remaining relatively stable in the preceding and following stages. The curve of the synthesized video is slightly higher than that of the real video, indicating that the inter-frame variation amplitude is slightly larger, and there are still slight tremors locally, but no obvious abnormal mutations have occurred.

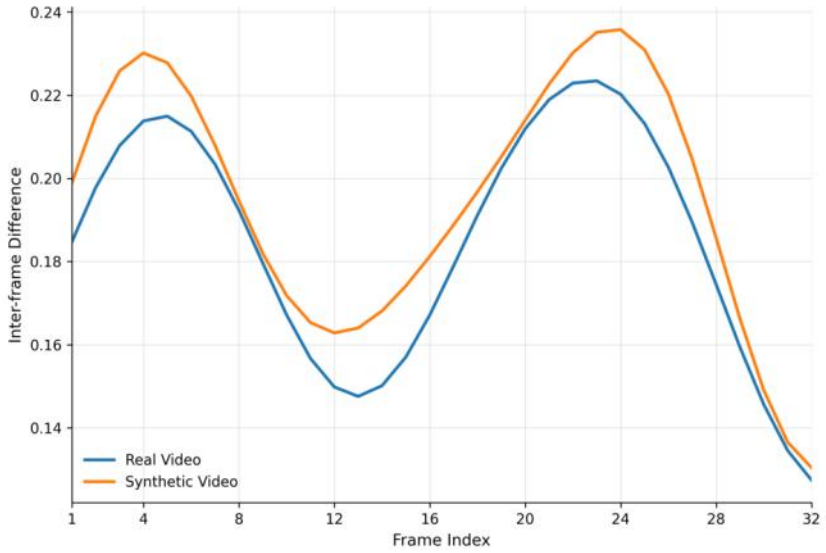

Figure 4: Curve graph showing the continuity changes between synthetic videos and real videos frames

From different action types, there are also certain differences in the quality of synthetic videos among various action categories. For projects with clearer action paths and more stable rhythm changes, the generated effect is usually better; while for rotation-type, jumping-type or actions with rapid transitions, the score gap is relatively more obvious, indicating that the model still has room for improvement in the stability of complex actions.As shown in Figure 5, regardless of the throwing actions of sports such as gymnastics, basketball, and volleyball, which have good action integrity and are easily recognizable in teaching, the difference between the two is very small. This indicates that the generation of such actions not only ensures integrity but also facilitates students' cognitive understanding. However, the scores for rotational and jumping actions are relatively low. Among them, the educational recognition rate indicates that this model still needs to be strengthened in reproducing the rhythm of difficult movements and narrating small sections; the arm swinging and basic gymnastic movements in running and jumping belong to the medium difficulty level, indicating that this type of model can complete medium difficulty movements relatively stably.

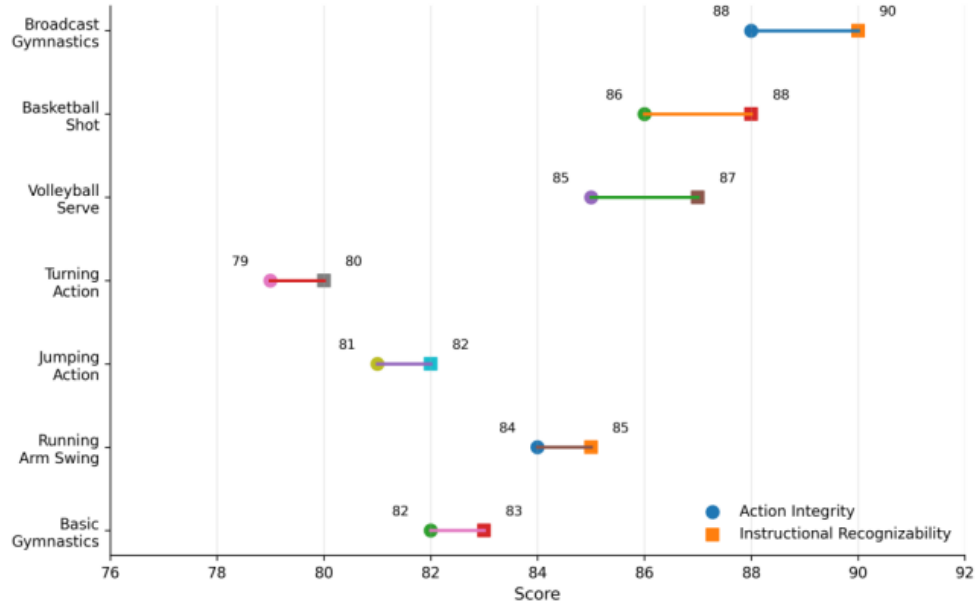

Figure 5: Comparison Chart of Completeness and Teaching Identifiability Scores for Different Action Types

To visually demonstrate the overall performance of each action category, this paper presents a heat map matrix for quality evaluation to reflect the correlation information between action categories and quality indicators, facilitating the analysis of the strengths and weaknesses of the model. As shown in Figure 6, the gymnastic routine has performed well in terms of structural similarity, temporal continuity, and teaching identifiability, indicating that the model has a stronger adaptability to projects with stable rhythms and repetitive movements; basketball shooting and volleyball serving also maintain a relatively high level overall, suggesting that the model has a better reconstruction ability in single-impact actions. In contrast, the rotational movements and jumping movements have relatively lower scores in posture reconstruction accuracy and temporal continuity, indicating that complex movements are more prone to errors in spatial changes and time connections.

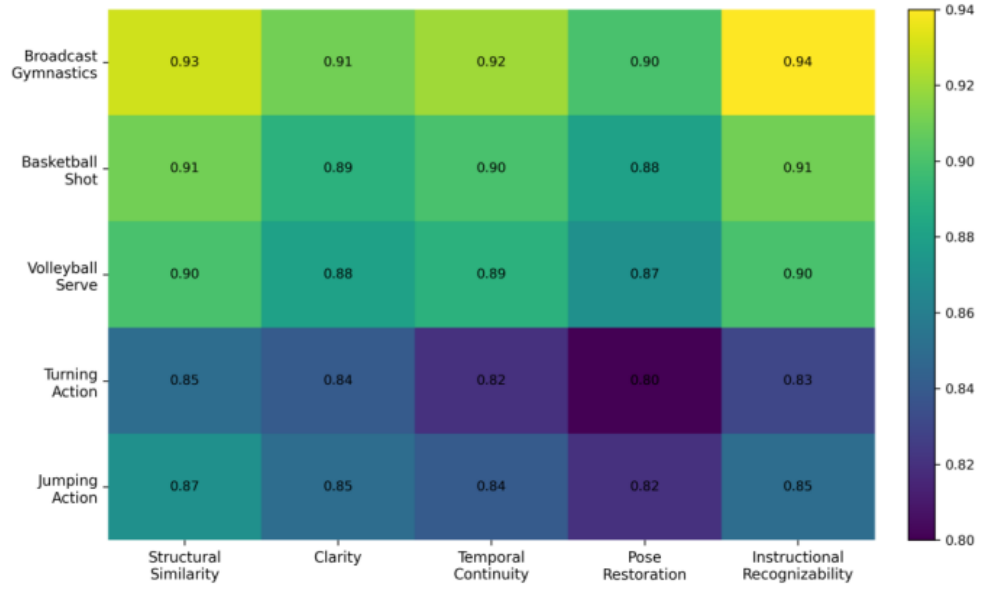

Figure 6: Heat map of video quality assessment for different action categories

From the results, it can be seen that the high-fidelity sports teaching demonstration videos generated in this study perform well in terms of overall visual quality, action continuity, and teaching recognizability. Although there are still certain deviations at the joints with rapid changes such as the wrist, knee, and ankle, and there are also certain fluctuations in complex actions such as rotation and jumping, from the overall results, the synthesized videos can already present the main structure and technical process of sports actions relatively stably, possessing a good foundation for teaching assistance applications, and providing reliable video materials for the evaluation of the effect of subsequent improvement in action recognition levels.

5.2 Evaluation of the effect of improving action recognition levels

To test the actual role of the high-fidelity sports teaching demonstration videos in teaching, this study evaluated the changes in students' action recognition levels in four aspects: understanding of action structure, recognition of action sequence, judgment of action errors, and memorization of key points. In the experiment, the experimental group and the control group were set up, and both groups had similar basic levels in the pre-test stage. Subsequently, they watched the synthesized demonstration videos and the conventional demonstration videos respectively, and completed the post-test and delayed test under the same teaching content and practice duration conditions. By comparing the test results at different stages, the influence of the synthesized demonstration videos on the improvement of action recognition can be judged more intuitively.

From the overall score changes, both groups of students showed certain improvements after the teaching intervention, but the improvement of the experimental group was more significant. In the pre-test stage, the total score of action recognition of the two groups was not significantly different, indicating that the starting points were relatively close; in the post-test stage, the total score of the experimental group was significantly higher than that of the control group, indicating that the high-fidelity synthesized demonstration videos had a more positive impact on action observation, action understanding, and action memory. As shown in Figure 7, the distribution centers of the experimental group and the control group in the pre-test stage were relatively close, both concentrated around 72 points; after the teaching intervention, the score

distribution of both groups shifted overall, but the shift of the experimental group was greater, with the post-test mean reaching 86.9 points, while that of the control group was 78.6 points. The length of the improvement arrows of the two groups in the figure shows a significant difference, indicating that the synthesized demonstration videos have a more significant driving effect on the total score of action recognition.

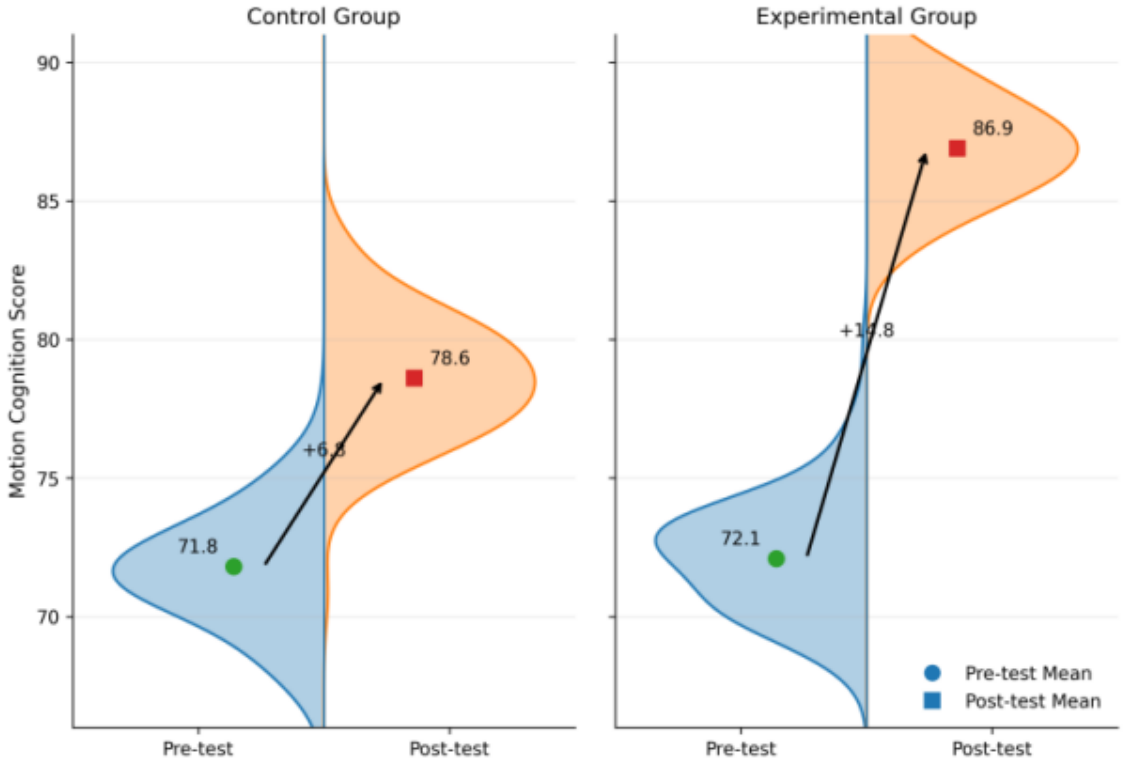

Figure 7: Comparison chart of the total scores of action cognition before and after the experiment between the experimental group and the control group.

From different cognitive dimensions, the experimental group outperformed the control group in four aspects: action structure understanding, action sequence recognition, action error judgment, and key point memory. Among them, the improvements in action sequence recognition and key point memory were the most significant, indicating that the synthesized demonstration videos have a significant advantage in presenting the action process and highlighting the key points; while the control group also showed a small improvement, but the differences in each indicator were not significant, and overall it was still lower than the experimental group. As can be seen from Figure 8, the median segments of the four items in the experimental group were all higher than those in the control group, and the maximum values were also better than those in the control group, indicating that the multi-level learning ability of the students in the experimental group has been significantly improved; especially for the two items of kinesthetic sequence recognition and key point memory, the high segments in the experimental group were more concentrated, indicating that high-quality demonstration videos can help students sort out the behavioral procedures and grasp the important details.

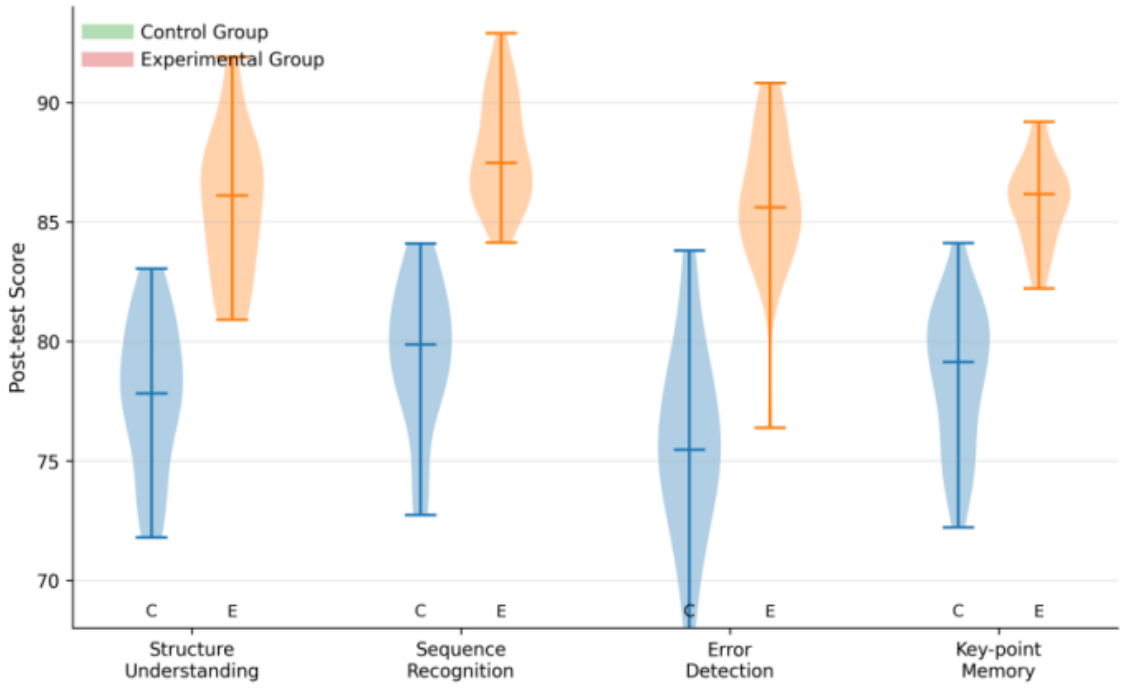

Figure 8: Violin plot of post-test scores for different cognitive dimensions

By comparing the various indicators of the experimental group, it was found that their development was more balanced and comprehensive. They showed significant improvements in their understanding of the action structure, memory ability, and comprehension level. This indicates that the synthesized videos help students establish a complete visual representation of the actions; they can better distinguish between incorrect and correct parts of the actions. It also shows that after watching high-quality demonstrations, students can better differentiate the differences between these two aspects. In contrast, the improvement range of the control group was overall smaller, especially in the aspects of error judgment and structure understanding, with the increase not as obvious as that of the experimental group. As shown in Figure 9, the improvement rate of the experimental group in all four dimensions was significantly higher than that of the control group. Among them, the increase in action sequence recognition and key point memory was the most prominent, reaching 20.3% and 19.4% respectively, while the improvement range of the corresponding indicators in the control group did not exceed 10%. This result indicates that the synthesized demonstration video not only enhances students' understanding of the overall structure of the actions, but also strengthens their cognition of the action process and key nodes.

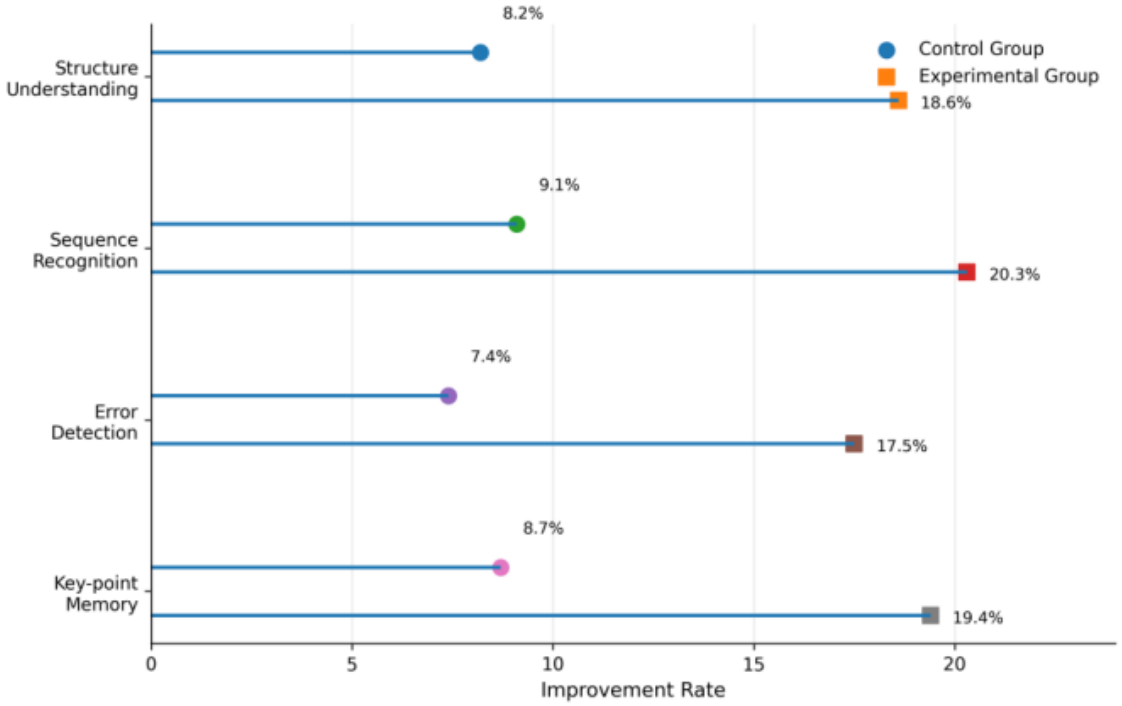

Figure 9: Comparison Chart of Action Cognition Improvement Rates between the Experimental Group and the Control Group

To further determine whether the teaching intervention effect has a certain degree of retention, this article arranged a delayed test after the post-test to track the students' retention of action cognition. From the results, both groups' scores would decline over time, but the decline in the experimental group was relatively smaller, and the overall retention level was higher than that of the control group. This indicates that the high-fidelity sports teaching demonstration videos not only can improve students' action cognition scores in a short period of time, but also help enhance students' continuous memory of the action structure and technical key points. As shown in Figure 10, the scores of the experimental group were always higher than those of the control group at the post-test, one week later, two weeks later, and four weeks later, and the decline curves were more gentle. The experimental group's score decreased from 86.9 points to 84.4 points, a decrease of 2.5 points; the control group's score decreased from 78.6 points to 75.5 points, a decrease of 3.1 points. Although both groups have a memory decay phenomenon, the retention effect of the experimental group is significantly better.

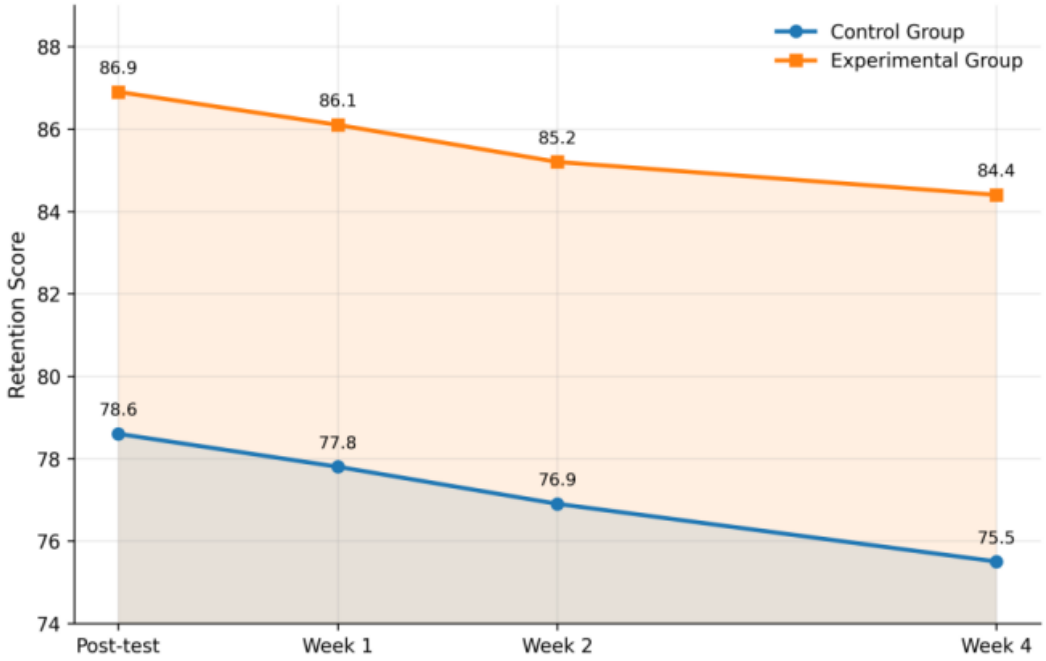

Figure 10: shows the changes in the effect of action cognition retention between the experimental group and the control group.

From the results, it can be seen that the high-fidelity sports teaching demonstration videos have a significant positive effect on improving action cognition. Compared with the conventional demonstration videos, the synthesized demonstration videos have more advantages in presenting the action structure, demonstrating the rhythm, and reinforcing key details, which can help students establish action imagery more quickly, improve the efficiency of action understanding, and enhance the retention of action key points. Especially in the aspects of action sequence recognition and key point memory, the performance of the experimental group was more outstanding. This indicates that high-quality generated demonstration videos not only enhance students' observation of action details, but also help them clarify the sequence of actions and grasp the technical points. Overall, such videos have strong practical application significance when used in physical education teaching.

5.3 Discussion

The research findings indicate that high-quality instructional demonstration videos for physical education have a relatively significant promoting effect on enhancing students' understanding of movements. Compared with traditional video teaching, the experimental group had a better understanding of the action structure, the ability to distinguish the sequence of actions, the ability to judge action errors, and a higher degree of memory for the main steps. This indicates that the synthetic video teaching can more effectively simplify the amount of information conveyed by the actions and promote students' grasp of the main information of the actions. This reflects the correlation between video quality and teaching effectiveness. If the movement process can be clearly displayed, with continuous changes and highlighting key points, it will be easier to establish a complete action image. In this study, the synthetic video has the advantages of coherent action structure, temporal continuity, and teaching intuitiveness, which may reduce students' kinesthetic dispersion rate and technical misinterpretation rate. However, the effects of different action categories are not all the same. For example, for rhythmic gymnastics, shooting, and throwing actions with strong complexity, the effect of synthetic

video is better; but for more complex actions, the above phenomenon changes to some extent. For example, for rotational jump actions, partial differences and time jitter still exist. This indicates that the current model is more suitable for basic sports action teaching, and there is still room for improvement in the generation of complex actions. From the perspective of teaching application, synthesized demonstration videos are more suitable as supplementary materials for teacher demonstrations. The advantage of teacher demonstrations lies in the ability to give on-site explanations and corrections, while the advantage of synthesized videos lies in the standardness of the actions, the ability to be repeatedly watched, and the convenience of review after class. The combination of the two is more conducive to improving the effectiveness of sports teaching. Overall, high-fidelity sports teaching demonstration videos based on generative adversarial networks have a promising application prospect. These videos help students establish clearer action cognition and provide a new implementation path for the digitalization and intelligence of subsequent sports teaching resources.

6 Conclusion

This paper focuses on the generation of high-fidelity sports teaching demonstration videos and the improvement of action cognition levels. A sports teaching demonstration video generation model based on generative adversarial networks was constructed, and its teaching application effects were verified. The research was analyzed from two aspects: video generation quality and action cognition improvement. It not only paid attention to the visual performance and action restoration of the synthesized demonstration videos, but also focused on their practical usage effects in sports teaching. The results showed that the synthesized videos achieved scores of 0.92, 0.90, 0.89, and 0.88 in terms of structural similarity, clarity, temporal continuity, and pose restoration, respectively, with overall good quality. The teaching experiment demonstrated that the total score of action cognition in the experimental group increased from 72.1 points to 86.9 points, while that in the control group increased from 71.8 points to 78.6 points. The improvement rates in action sequence recognition and key point memory of the experimental group reached 20.3% and 19.4% respectively, both higher than those of the control group. In the delayed test, the scores of the experimental group decreased from 86.9 points to 84.4 points, while those of the control group decreased from 78.6 points to 75.5 points. The retention effect of the experimental group was more stable. It indicates that high-fidelity sports teaching demonstration videos can effectively promote students' understanding of the action process and technical points, and have certain teaching application value.

About the Author

Jun Cao was born inJiangxi, in 1979. Graduate student, Master's degree holder, Lecturer, Associate Dean of the School of Physical Education. Served as Vice Principal of Yuzhang Primary School in Xihu District, Nanchang City from 2017 to 2019. Core faculty member of the School of Physical Education. National Level 1 Basketball Referee and National Level 1 Amateur Basketball Coach. Led 5 provincial-level research projects, participated in 6 provincial-level and 2 university-level projects, and co-authored 4 textbooks. Taught courses including Basketball, Volleyball, Badminton, Gymnastics, Elementary Physical Education Teaching and Curriculum Theory, and College Student Physical Education and Health. Honored as National Outstanding Coach and Jiangxi Province Outstanding College Basketball Referee. caojun7211@126.com

References

[1] Yang Liu.(2025).Research on the Application of Micro-video + Comprehension Teaching Method in College Physical Education Courses——Taking Basketball as an Example. International Educational Research Development, 2(4), 23-25. https://doi.org/10.12462/IERD.ISSN3007-7664.2025.04.008.

[2] Omar Trabelsi,Amir Romdhani,Ahmed Ghorbel,Mustapha Bouchiba,Mohamed Abdelkader Souissi,Swantje Scharenberg & Adnene Gharbi.(2025).A review of best practices in video modeling for sport pedagogues.International Journal of Sports Science & Coaching,20(4),1749-1760.https://doi.org/10.1177/17479541251335611.

[3] Montassar Ben Romdhane,Hajer Mguidich,Houssem Ben Chikha,Hamdi Chtourou & Aïmen Khacharem.(2025).Optimizing Basketball Tactics Learning in Physical Education: The Impact of Modality and Video Control..Perceptual and motor skills, 132(5), 315125251328727.https://doi.org/10.1177/00315125251328727.

[4] Alex Adams, Tyler Goad, Alysia Jenkins & Don Belcher. (2025). Powering Up Online Physical Education: Unleashing the Potential of Video. Strategies, 38(2), 13-20. https://doi.org/10.1080/08924562.2024.2444205.

[5] Alex Adams,Don Belcher,Alysia Jenkins & Tyler Goad.(2025).Utilizing Video Feedback for Effective Motor Skill Learning in Online Physical Education.International Journal of Kinesiology in Higher Education, 9(1), 14-30. https://doi.org/10.1080/24711616.2024.2421764.

[6] Valérian Cece, Patrick Fargier, Cédric Roure & Vanessa Lentillon Kaestner.(2025). Multidisciplinary teaching with an active video game: the effect on learning in mathematics and physical education.Technology, Pedagogy and Education, 34(1), 121-136. https://doi.org/10.1080/1475939X.2024.2407380.

[7] Zou Yan & Han Yan.(2024).Research on the Application Methods of Short Video APP in Physical Education Teaching in Colleges and Universities.Frontiers in Sport Research, 6(6), https://doi.org/10.25236/FSR.2024.060612.

[8] Xiangyang Wang. (2024). Application of Interactive Object Model in Sports Teaching. International Journal of High Speed Electronics and Systems, 34(02), https://doi.org/10.1142/S0129156424400871.

[9] Haipeng Wan, Xue Zhang, Xinxue Yang & Shan Li.(2024).Which approach is effective: Comparing problematization-oriented and structuring-oriented scaffolding in instructional videos for programming education.Education and Information Technologies, 29(14),17807-17823.https://doi.org/10.1007/S10639-024-12550-0.

[10] Applied Bionics And Biomechanics.(2024). Retracted: Image Video Teaching Method in College Physical Education..Applied bionics and biomechanics, 2024, 9896734-9896734. https://doi.org/10.1155/2024/9896734.

[11] Ling Cao,Gang Hao & Houmin Wu.(2024).AIGC enables vocational education teaching-take the course of Audio and video editing processing as an example.Applied

Mathematics and Nonlinear Sciences,9(1),https://doi.org/10.2478/AMNS-2024-3008.

[12] Yangsheng Zhang,Peijun Wei & Michelle Ranges.(2024).An Innovative Approach to a Physical Education Skills Teaching System Incorporating Online Video.Applied Mathematics and Nonlinear Sciences,9(1),https://doi.org/10.2478/AMNS-2024-3274.

[13] Jinming Liu,Sheng Yang,Dunlin Zhu,Tianyun Luo,Jinglong He & Shengyuan Li.(2025).Intelligent error-proof logic multi-factor cooperative active optimization based on knowledge base and generative adversarial network.Systems Science & Control Engineering,13(1),https://doi.org/10.1080/21642583.2025.2486126.

[14] Jihwan Shin,Yeji Song,Minsoo Jang,Jinhyun Ahn,Taewhi Lee & Dong Hyuk Im.(2025). Differential privacy in statistical queries for synthetic trajectories generated by generative adversarial networks.Connection Science, 37(1), https://doi.org/10.1080/09540091.2025.2523964.

[15] Yuyang Wang & Qiaowei Xue.(2025).Retraction Note: Fault identification of product design using fuzzy clustering generative adversarial network (FCGAN) model.Soft Computing,(prepublish),1-1.https://doi.org/10.1007/S00500-025-11078-W.

[16] Mars Caroline Wibowo, Daniel Manongga, Hendry Hendry & Teguh Indra Bayu.(2025). Enhancing character animation realism with generative adversarial networks (GANs): a comparative method study. Discover Artificial Intelligence, 5(1), 398-398. https://doi.org/10.1007/S44163-025-00501-8.

[17] Marjan Ilbeigi, Mohammad Ghomeishi, Ali Asgharzadeh & Nima Amani.(2025). Integrating post occupancy evaluation and energy performance metrics using conditional generative adversarial networks for intelligent school design.Discover Sustainability, 7(1), 153-153. https://doi.org/10.1007/S43621-025-02523-9.

[18] Zhenghua Xu, Jiaqi Tang, Dan Yao, Zhenzhen Wang & Thomas Lukasiewicz. (2026). AttCL-GAN: Attentional contrastive learning-based generative adversarial network for modality completion of medical images. Knowledge-Based Systems, 334, 115017-115017. https://doi.org/10.1016/J.KNOSYS.2025.115017.

[19] Samantha J. Brozak, David Montes de Oca Zapiain,Brendan Donohoe,Tommy Ao,Nathan P. Brown, Marcus D. Knudson... & J. Matthew D. Lane.(2026).Machine learning for domain transfer between simulated and experimental 2D X-ray diffraction patterns using generative adversarial networks.Computational Materials Science, 263, 114429-114429.https://doi.org/10.1016/J.COMMATSCI.2025.114429.

[20] Yuelin Yang, Chunjie Yang, Siwei Lou, Dali Gao, Xujie Zhang & Weibin Wang.(2026). CSNF-TimeGAN: Class-prior spatiotemporal nonlinear feature-based time series generative adversarial network for blast furnace fault diagnosis.Advanced Engineering Informatics,71(PA),104213-104213.https://doi.org/10.1016/J.AEI.2025.104213.

[21] Han Qi,Yihan Xu,Hao Wang,Abdullah Gani & Lip Yee Por.(2025).HQWGAN: a hybrid quantum–classical Wasserstein generative adversarial network for image generation.The Journal of Supercomputing,82(1),25-25.https://doi.org/10.1007/S11227-025-08172-Z.