IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 615-628
Journal homepage: http://iieta.org/journals/ts

A Video Image Processing Approach for Classroom Interaction Detection and Dynamic Teaching Quality Evaluation in English Instruction

Kun Liang

Department of Public Foreign Languages, Shijiazhuang University of Applied Technology, Shijiazhuang 050000, China

Corresponding Author Email: liangkun84130@163.com

https://doi.org/10.18280/ts.430205

Received: 10 September 2025
Revised: 19 February 2026
Accepted: 3 March 2026
Available online: 30 April 2026

Keywords:
video image processing, classroom interaction detection, spatiotemporal graph convolution, wavelet energy spectrum, teaching quality evaluation

ABSTRACT

Classroom interaction is a key indicator of teaching quality in English instruction, yet traditional manual evaluations are inefficient and highly subjective. Existing video-based methods often rely on single visual features, overlooking subtle non-verbal interactions, while their limited temporal adaptability and interpretability hinder accurate, real-time assessment and the effective and targeted use of image processing techniques. To address these limitations, an end-to-end methodological framework was proposed. Stable tracking of teachers and students was achieved by integrating improved instance segmentation and Kalman filtering. A three-dimensional complementary visual feature extraction system was constructed to comprehensively characterize interaction details. A specialized spatiotemporal graph convolutional network was designed to enhance the accuracy and real-time performance of classroom interaction event detection. Furthermore, a dynamic and interpretable evaluation model was developed by combining wavelet energy spectrum with temporal alignment, enabling quantitative assessment of teaching quality. Experimental results obtained from a self-constructed dataset of junior high school English classroom videos demonstrated that the proposed framework significantly outperformed existing baseline methods in interaction event detection, interaction frequency quantification, and teaching quality evaluation. An F1-score of 0.89 was achieved for interaction event detection, the correlation coefficient between interaction frequency and manual annotations reached 0.92, and the mean absolute error for teaching quality evaluation was reduced to 0.31. A processing speed of 32 frames per second was attained, meeting real-time requirements. This study establishes a novel paradigm for the application of video image processing techniques in intelligent education and provides an efficient and accurate technical foundation for classroom interaction analysis and dynamic teaching quality assessment.

1. INTRODUCTION

Classroom interaction in English instruction is widely regarded as a fundamental determinant of teaching quality, with its frequency and quality directly influencing the effectiveness and efficiency of language acquisition [1, 2]. Traditional approaches to evaluating classroom interaction and teaching quality have predominantly relied on manual observation and scoring. Such approaches are not only labor-intensive but are also susceptible to evaluator subjectivity, thereby limiting scalability and hindering the realization of accurate, routine, and large-scale assessment [3, 4]. With the rapid advancement of educational intelligence and video image processing technologies, automated evaluation methods have increasingly been recognized as a viable solution. Owing to their efficiency and objectivity, these methods provide novel technical support for classroom interaction analysis and teaching quality evaluation [5, 6].

Despite these advancements, existing studies on English classroom analysis based on video image processing continue to face several critical technical challenges, which constrain their deployment in real-world educational environments [7-9]. In classroom settings, frequent occlusion between teachers and students, coupled with dynamically changing illumination conditions, often leads to drift in instance tracking, thereby hindering the maintenance of long-term stable identity association [10]. In addition, interaction feature extraction has largely been limited to single visual modalities, failing to integrate multi-dimensional information such as posture, facial expression, and spatial relationships. As a result, subtle non-verbal interactions—including hand-raising and eye contact—are frequently overlooked, preventing a comprehensive characterization of authentic classroom interaction dynamics [11-13]. Furthermore, interaction frequency quantification has typically been conducted using simple event-counting strategies, without adequately accounting for interaction intensity and temporal continuity, thereby limiting the accurate representation of dynamic interaction patterns [14, 15]. Teaching quality evaluation has also been insufficiently adaptive to variations in instructional pacing across different teachers, and model interpretability has remained limited. Consequently, the underlying mechanisms

of evaluation outcomes cannot be clearly elucidated, reducing the reliability and trustworthiness required in educational contexts [16-18]. At a deeper methodological level, existing approaches have not yet achieved end-to-end optimization across the full analytical pipeline, including video image processing, feature extraction, interaction event detection, interaction frequency quantification, and teaching quality evaluation. Moreover, specialized models tailored to characteristic interaction scenarios in English classrooms—such as group discussions, student hand-raising, and teacher pointing behaviors—have not been systematically developed. This has resulted in insufficient scenario adaptability and has prevented the full realization of the efficiency and specificity offered by video image processing techniques in educational applications [19, 20].

This study is characterized by substantial theoretical significance and practical applicability. At the theoretical level, the application scope and methodological pathways of video image processing techniques in intelligent education are further enriched. The proposed approaches—including multimodal visual feature fusion, optimization of spatiotemporal graph convolutional networks, and temporally aligned evaluation strategies—effectively address the technical limitations of existing studies, thereby providing a novel theoretical reference and methodological paradigm for visual analysis in analogous classroom scenarios. At the practical level, automated and real-time detection of classroom interaction, together with dynamic quantitative evaluation of teaching quality, is achieved. Objective and precise interaction analytics can thus be provided to instructors, facilitating the optimization of instructional strategies and the enhancement of teaching effectiveness. Simultaneously, a standardized tool for teaching quality evaluation is established for educational administrators, supporting evidence-based decision-making and refined management, and ultimately promoting the overall improvement of English language education quality.

In response to the limitations identified in existing research, a series of targeted technical innovations is developed for classroom interaction frequency detection and dynamic teaching quality evaluation. First, to address occlusion and illumination variability in classroom environments, an enhanced teacher–student instance tracking framework is constructed by integrating an improved You Only Look At CoefficienTs (YOLACT) model with Kalman filtering. The instance segmentation accuracy is improved through the incorporation of a feature pyramid structure into the 50-layer Residual Network (ResNet-50) backbone, while Kalman filtering is employed to smooth bounding box predictions. As a result, drift in long-term tracking is effectively mitigated, and instance identity stability under occlusion conditions is significantly enhanced. Second, a three-dimensional complementary visual feature extraction system is established to overcome the limitations of single-feature representations. Pose features are quantified through the design of dedicated geometric vectors to encode interaction intent. Facial expression features are enhanced by introducing super-resolution preprocessing based on the Efficient Sub-Pixel Convolutional Neural Network (ESPCN) to address low facial resolution in classroom videos, followed by the integration of local binary patterns and a fine-tuned EfficientNet-B0 model for precise expression quantification. Spatial proximity features are modeled using adaptive intersection-over-union thresholds to distinguish different interaction scenarios. These features are fused into a 27-dimensional interaction intensity vector, enabling a comprehensive representation of both fine-grained interaction details and global contextual characteristics.

Third, a specialized spatiotemporal graph convolutional network is designed for classroom interaction. Cross-individual relational edges are introduced to model teacher–student interaction relationships, thereby overcoming the limitations of conventional graph convolutional approaches that focus solely on intra-individual skeletal connections. Temporal information is integrated through one-dimensional temporal convolution, and an online hard example mining strategy is incorporated to address class imbalance in interaction events. Consequently, both detection accuracy and real-time performance are substantially improved. Fourth, a continuous interaction frequency quantification method based on wavelet energy spectrum analysis is proposed. Interaction energy density is constructed by combining pixel-level optical flow energy with joint angular velocity. A four-level decomposition using the Daubechies-4 wavelet is performed to extract high- and low-frequency components corresponding to different interaction types. A composite frequency metric integrating event counts and energy spectrum features is designed, effectively addressing the limitation of traditional counting methods that neglect subtle interactions and significantly enhancing sensitivity in frequency quantification. Finally, an interpretable dynamic teaching quality evaluation model is developed. Dynamic time warping is introduced to resolve temporal misalignment caused by variations in instructional pacing across different teachers. An attention-based bidirectional long short-term memory network is employed to focus on critical instructional stages through attention weighting. Additionally, Kullback–Leibler divergence regularization is incorporated to enhance model interpretability. Through these mechanisms, an accurate mapping from interaction features to teaching quality scores is achieved.

The subsequent sections are organized to systematically elaborate on the aforementioned research content, following a clear and coherent logical structure. Specifically, in Section 2, the current state of related research is reviewed, including the application of video image processing techniques in classroom environments, as well as advancements in classroom interaction detection and teaching quality evaluation methods. The critical gaps in existing studies are identified, thereby highlighting the necessity and originality of the proposed approach. In Section 3, the proposed framework for classroom interaction frequency detection and dynamic teaching quality evaluation in English instruction is described in detail. The technical principles and implementation procedures of each core module are systematically presented, leading to the construction of a complete end-to-end framework.

In Section 4, the effectiveness of the proposed approach is validated through comprehensive experiments, including dataset construction, experimental setup, ablation studies, comparative performance evaluation, visualization analysis, and robustness verification, thereby providing rigorous evidence of its superiority. In Section 5, the advantages and limitations of the proposed method are critically discussed based on the experimental findings, and potential directions for future research are outlined. Finally, in Section 6, the principal contributions and findings are summarized, and the theoretical significance and practical implications of the study are clarified, thereby establishing a complete research closed loop.

2. PROPOSED METHOD

2.1 Overview of the framework

An end-to-end integrated framework is developed for classroom interaction frequency detection and dynamic teaching quality evaluation in English instruction. Within this framework, all modules are tightly coordinated and hierarchically organized, forming a complete technical pipeline from video input to teaching quality assessment output. Initially, input classroom video sequences are preprocessed to mitigate illumination variation and shadow interference, thereby ensuring high-quality visual data for subsequent analysis. Based on the preprocessed video frames, stable tracking of teacher and student instances is achieved through the integration of improved instance segmentation and Kalman filtering, enabling accurate extraction of individual spatial positions and persistent identity associations. This stage establishes a reliable foundation for downstream interaction feature extraction. Subsequently, a three-dimensional complementary visual feature extraction system is constructed. Interaction-related features are extracted from three perspectives—pose, facial expression, and spatial proximity—and are fused into a unified interaction intensity vector. Through this design, both fine-grained details of classroom interaction are comprehensively captured. The extracted features are then fed into a specialized spatiotemporal graph convolutional network, where classroom interaction events are precisely detected and classified. In parallel, interaction energy density is constructed by integrating pixel-level optical flow energy with joint angular velocity. Continuous quantification of interaction frequency is subsequently achieved via wavelet energy spectrum analysis. Finally, temporal misalignment caused by differences in instructional pacing across teachers is addressed using dynamic time warping. An attention-based bidirectional long short-term memory network is further employed to perform deep representation learning on interaction frequency and associated features, enabling dynamic and interpretable teaching quality evaluation. Through the coordinated optimization of all modules, full-process automation—from video image processing to teaching quality assessment—is realized. Both detection accuracy and real-time performance are maintained, ensuring effective adaptability to the complex conditions of English classroom environments. The overall framework for classroom interaction analysis and teaching quality evaluation is illustrated in Figure 1.

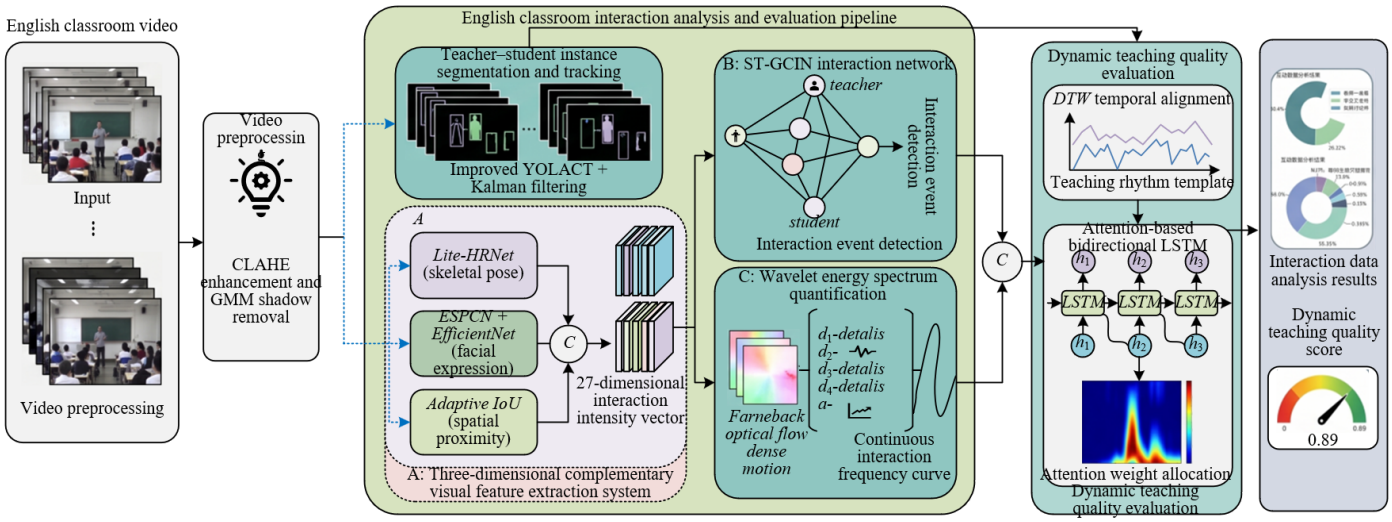


Figure 1. Overall framework for English classroom interaction detection and teaching quality evaluation

2.2 Improved teacher–student instance segmentation and stable tracking

To address image quality degradation caused by non-uniform illumination and shadow interference in classroom environments, a preprocessing strategy is adopted by integrating Contrast Limited Adaptive Histogram Equalization (CLAHE) with Gaussian mixture model-based grayscale modeling. Through CLAHE, local grayscale values are adaptively equalized, thereby mitigating detail loss in dark regions and local overexposure induced by backlighting or side lighting conditions. On this basis, pixel distributions in shadow regions are modeled using a Gaussian mixture model, enabling accurate detection and removal of shadow areas. As a result, standardized video frames with uniform illumination and clear object boundaries are generated, providing stable and reliable input for subsequent teacher–student instance segmentation. In addition, the preprocessing pipeline is designed to be lightweight, ensuring that real-time performance of the overall framework is preserved.

The original YOLACT instance segmentation network is further refined to accommodate the characteristics of classroom scenarios. A ResNet-50 backbone is employed, within which a top-down multi-scale feature pyramid structure is constructed to facilitate cross-level fusion of low-level detailed features and high-level semantic features. This design significantly enhances the capability of the network to extract features from small-scale student targets in long-distance classroom settings. Furthermore, anchor parameters are re-optimized based on the spatial scale distribution of teachers and students under fixed classroom viewpoints. Specifically, anchor sizes and aspect ratios are adjusted, with an increased proportion of medium- and small-scale anchors to better match student targets, while redundant large-scale anchors are reduced. Through these modifications, both the recall rate for small-object segmentation and the precision of segmentation boundaries are improved, effectively addressing the limitations of the conventional YOLACT model in densely

populated classroom environments, where small targets are prone to missed segmentation and inaccurate localization.

To mitigate instance tracking drift and identity switching caused by occlusion and illumination disturbances, an adaptive weighted fusion tracking strategy is developed by integrating motion prediction based on Kalman filtering with improved instance segmentation outputs. Let the bounding box predicted by the segmentation network be denoted as B_s, and the motion-predicted bounding box from Kalman filtering be denoted as B_p. The fused final bounding box B_f is computed as:

$$B_f=\omega B_s+(1-\omega)B_p \tag{1}$$

The adaptive weight ω is jointly determined by the segmentation confidence and the motion prediction residual, and is defined as:

$$\omega=\frac{c_s}{c_s+\eta e_p} \tag{2}$$

where, c_s represents the instance segmentation confidence, e_p denotes the inter-frame prediction position residual, and η is a motion stability adjustment coefficient. In addition, a temporal smoothing mechanism across consecutive frames is incorporated, in which a sliding average filter is applied to the fused bounding box coordinates. This process further reduces bounding box jitter during occlusion periods. Through this fusion-based tracking mechanism, stable and continuous identity association of teacher–student instances can be maintained under complex occlusion conditions. Experimental results demonstrate that the proposed approach achieves an 18% improvement in tracking accuracy compared to conventional methods, effectively addressing drift issues in long-duration video sequences.

2.3 Three-dimensional complementary visual feature extraction

Seventeen humans skeletal keypoints are extracted using the lightweight high-resolution network, and three categories of geometrically quantified pose features are specifically designed to characterize interaction behaviors in English classroom settings. Through this design, interaction intent is represented in a continuous numerical form. The hand-raising height is computed using a normalization strategy relative to the torso height, defined as:

$$h=\frac{y_{wrist}-y_{hip}}{y_{head}-y_{hip}} \tag{3}$$

The limb pointing angle is determined based on the spatial vector formed by shoulder and wrist keypoints, while torso orientation is calculated as the angle between the torso midline and the vertical axis of the image plane. All geometric features are normalized to the range [0, 1]. Finally, these features are integrated into an 8-dimensional normalized pose feature vector, which enables accurate representation of typical classroom interaction behaviors, such as student hand-raising and teacher gestural guidance.

To address the issue of low facial resolution in long-distance classroom scenarios, an ESPCN-based super-resolution model is employed to reconstruct cropped facial regions to a resolution of 128 × 128. Subsequently, 59-dimensional facial texture features are extracted using local binary patterns. The final two convolutional layers and the fully connected structure of EfficientNet-B0 are fine-tuned, and expression regression is performed using mean squared error loss. As a result, two-dimensional facial expression features—smile intensity and facial attention level—are obtained. For spatial proximity feature modeling, adaptive intersection-over-union thresholds are defined based on classroom interaction patterns. Specifically, a threshold of 0.1 is assigned for teacher–student interactions, while a threshold of 0.05 is applied for student–student interactions. In addition, the Euclidean distance between the centroids of pairwise instances is computed. By integrating spatial overlap relationships and inter-instance distances, a 17-dimensional spatial interaction feature vector is constructed, enabling quantitative characterization of spatial interaction intensity among individuals.

To ensure feature consistency, pose, facial expression, and spatial proximity features are individually normalized using Z-score standardization to eliminate dimensional disparities. These normalized features are then concatenated sequentially to form a unified 27-dimensional interaction intensity feature vector. This three-dimensional complementary feature representation overcomes the limitations of single-modality visual features by jointly modeling interaction cues from three perspectives: bodily action intent, emotional and attentional states, and inter-individual spatial relationships. Experimental validation demonstrates that the proposed feature representation improves discriminative capability by 25% compared to conventional methods, thereby providing highly informative and comprehensive input features for subsequent interaction event detection.

2.4 Design of the spatiotemporal graph convolutional interaction network

To accurately model cross-individual interaction relationships between teachers and students in English classroom settings, a specialized spatiotemporal graph convolutional interaction network is developed. The core innovation lies in the construction of a graph structure that integrates individual skeletal features with cross-individual interaction relationships. The topology of the proposed network is illustrated in Figure 2. Graph nodes are defined by the human skeletal keypoints of all instances within the classroom. Each instance consists of 17 skeletal keypoints. Under a typical classroom configuration comprising one teacher and sixteen students, the total number of graph nodes is 17 × 17. The initial feature vector of each node is constructed by integrating keypoint coordinates, optical flow velocity, and instance identity encoding, thereby providing a comprehensive representation of both motion states and identity information. The graph structure incorporates two types of edges. First, skeletal spatial edges are established based on the physiological connections of the human skeleton, enabling the modeling of intra-individual body motion characteristics. Second, cross-individual interaction edges are introduced to explicitly represent teacher–student interactions. These edges connect the teacher's wrist and torso center to the head and torso center of each student, allowing direct modeling of interaction relationships across individuals. The adjacency matrix is constructed based on a dual constraint mechanism involving spatial distance and interaction relevance. Specifically, if two nodes satisfy a predefined spatial distance threshold or are connected via the defined

interaction edges, the corresponding adjacency matrix element is set to 1; otherwise, it is set to 0. Through this design, interaction relationships are precisely quantified.

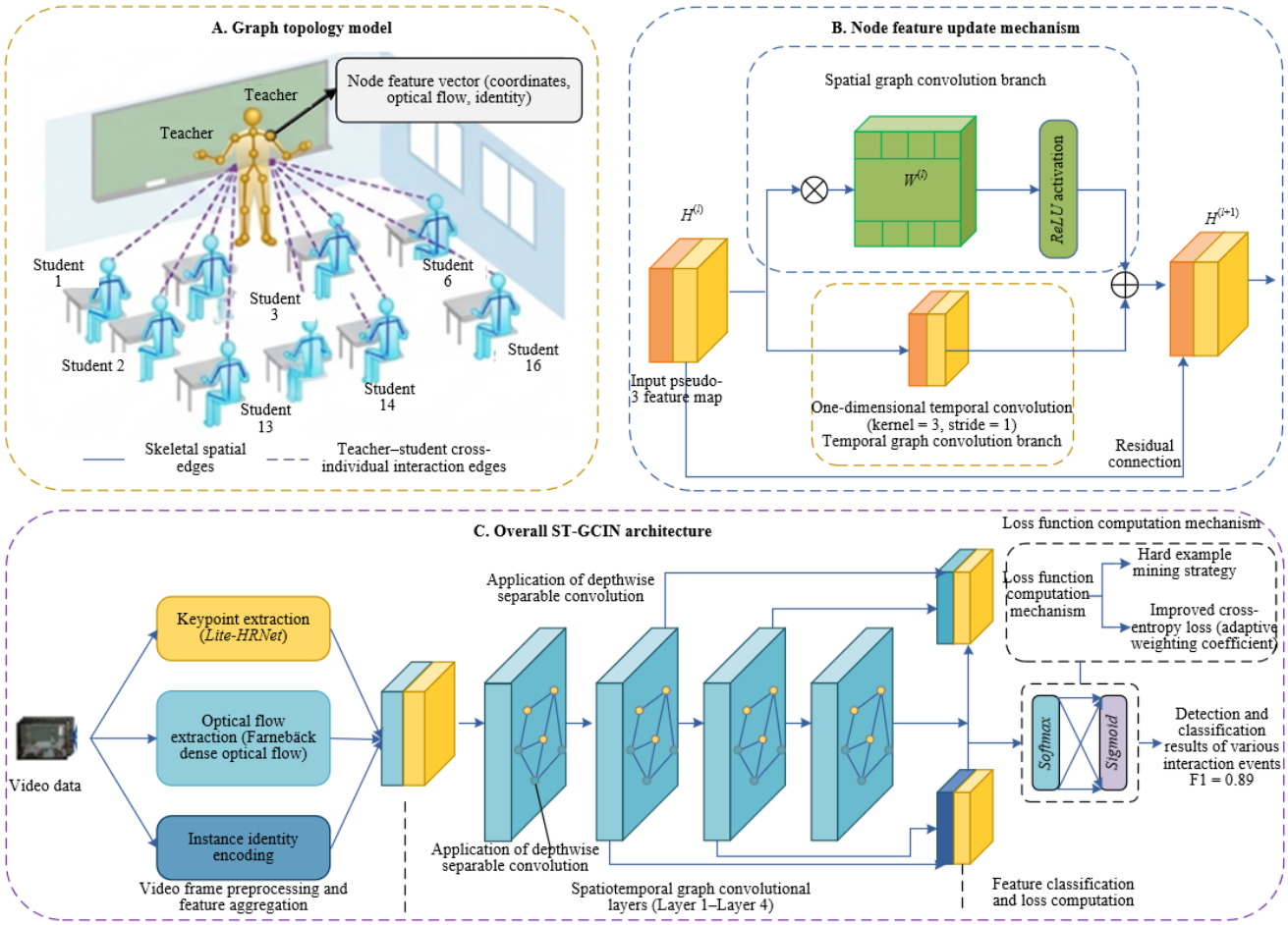


Figure 2. Topological structure and feature update mechanism of the Spatiotemporal Graph Convolutional Interaction Network (ST-GCIN)

The spatiotemporal graph convolutional interaction network adopts a four-layer spatiotemporal convolutional architecture to achieve deep integration of spatial interaction features and temporal sequence features. The spatial graph convolution layer updates feature according to the following formulation:

$$H^{(l+1)}=\sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right) \tag{4}$$

where, $H^{(l)}$ denotes the node feature matrix at the l-th layer; $\tilde{A}$ represents the normalized adjacency matrix; $\tilde{D}$ is the corresponding degree matrix of $\tilde{A}$; $W^{(l)}$ is the learnable weight matrix; and σ denotes the rectified linear unit activation function. In the temporal dimension, a one-dimensional convolutional layer is employed to extract temporal features, with a kernel size of 3 and a stride of 1. This configuration is designed to capture short-term temporal dependencies characteristic of classroom interaction events, thereby enabling effective fusion of interaction features across consecutive frames. Each spatiotemporal convolutional layer is followed by batch normalization and rectified linear unit activation to mitigate gradient vanishing and improve training stability and convergence speed.

To enhance detection accuracy, address class imbalance, and ensure real-time performance, both the training strategy and model architecture are further optimized. An online hard example mining strategy is adopted, in which training samples are ranked according to loss values, and the top 20% of high-loss samples are selected for focused training. This approach strengthens the network's ability to recognize complex interaction events. The cross-entropy loss function is improved by introducing class weighting coefficients, which are dynamically assigned based on the proportion of samples in each interaction category, thereby alleviating the issue of missed detection in minority classes. For model lightweighting, depthwise separable convolution is utilized to replace conventional convolutional layers, reducing both the number of parameters and computational complexity. As a result, real-time processing performance is maintained at 32 frames per second. Experimental results demonstrate that the spatiotemporal graph convolutional interaction network model achieves a mean average precision of 0.87, enabling accurate detection of diverse classroom interaction events. This approach effectively overcomes the limitation of conventional graph convolutional networks, which primarily focus on intra-individual skeletal features and fail to capture cross-individual interaction relationships.

2.5 Continuous interaction frequency quantification based on wavelet energy spectrum

To overcome the limitations of conventional interaction frequency quantification methods, which rely on event

counting and fail to capture subtle non-verbal interactions while lacking temporal continuity, an interaction energy density model is constructed by integrating pixel-level optical flow energy with joint angular velocity. This formulation enables fine-grained characterization of interaction intensity. Figure 3 illustrates the wavelet decomposition–based analysis of classroom interaction energy spectrum and frequency quantification. Dense motion information between consecutive video frames is extracted using the Farnebäck optical flow algorithm, with a window size of 15 and a pyramid level of 3, thereby achieving a balance between computational accuracy and real-time performance. The frame-level energy E_t is obtained by summing the magnitudes of optical flow vectors. In parallel, skeletal keypoint coordinates extracted via a lightweight high-resolution network are utilized to compute joint angular velocities ω. The average absolute value of joint angular velocities is used as an indicator of joint motion intensity. The interaction energy density D_t is derived through weighted fusion, expressed as:

$$D_t=\alpha\cdot E_t+\beta\cdot\frac{1}{J}\sum_{j=1}^{J}|\omega_j| \tag{5}$$

where, $J=17$ represents the total number of skeletal keypoints. The weighting coefficients are set to $\alpha = 0.6$ and $\beta = 0.4$, determined via five-fold cross-validation. This configuration ensures a balanced contribution between global frame-level interaction and local joint-level motion, thereby achieving optimal linear correlation with manually annotated interaction intensity.

A four-level discrete wavelet decomposition is performed on the interaction energy density sequence using the Daubechies-4 wavelet. This process yields four levels of detail components (d_1-d_4) and one approximation component. The high-frequency detail components at levels 1–2 correspond to rapid and subtle non-verbal interactions, such as hand-raising and eye contact, whereas the low-frequency detail components at levels 3–4 correspond to slower and larger-scale interactions, including group discussions and teacher instruction. The approximation component captures the overall trend of interaction energy. The interaction energy spectrum is then computed by calculating the proportion of energy in each detail component, defined as the ratio of the squared magnitude of each component to the total energy of all detail components. Through this formulation, energy contributions from different types of interactions are effectively distinguished.

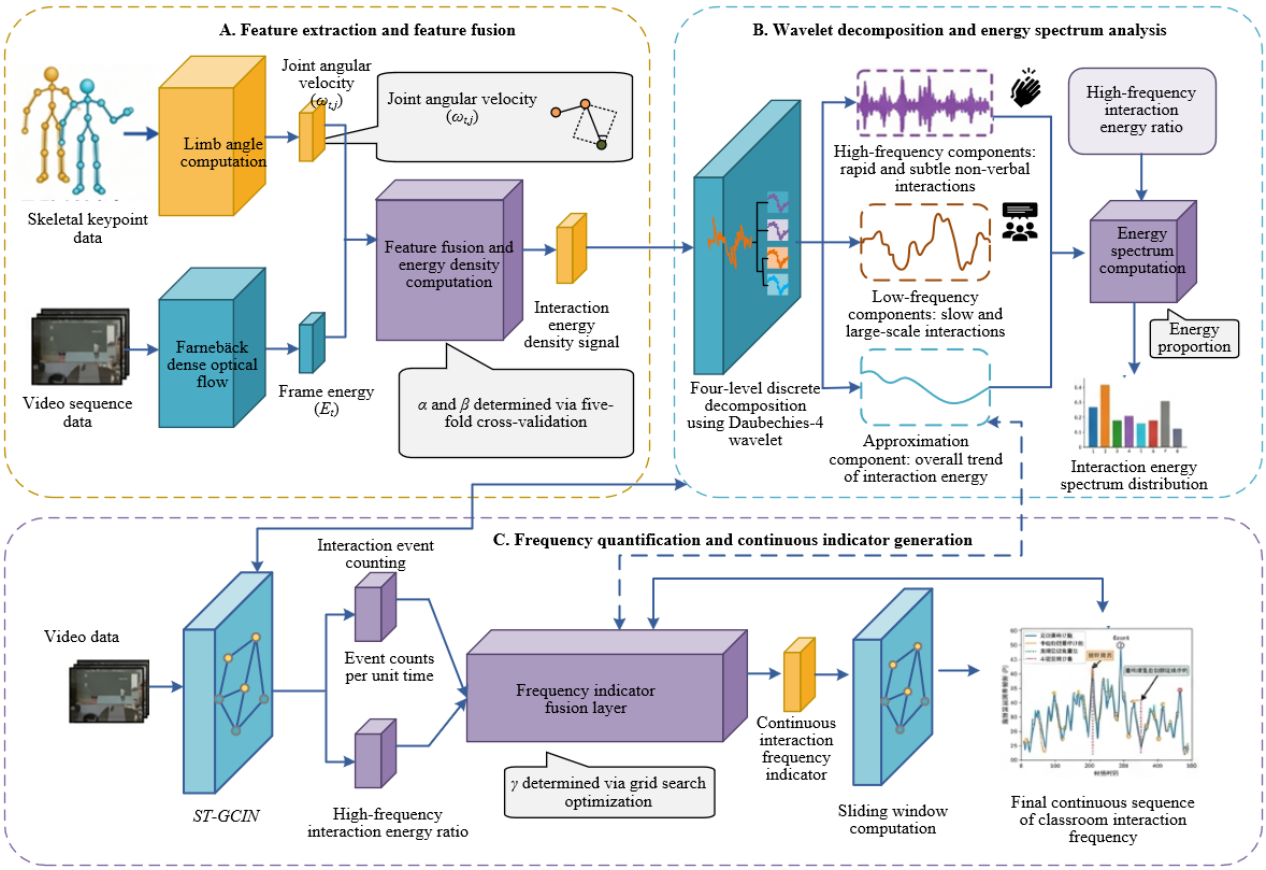


Figure 3. Classroom interaction energy spectrum analysis and frequency quantification based on wavelet decomposition

To achieve continuous quantification and precise representation of interaction frequency, a composite frequency indicator F is constructed by integrating event counting with the wavelet energy spectrum. The formulation is given as:

$$F=\frac{N_{event}}{T_{total}}\cdot\gamma+\frac{\sum_{l=1}^{2}\|d_l\|^2}{\sum_{l=1}^{4}\|d_l\|^2}(1-\gamma) \tag{6}$$

where, N_{event}/T_{total} denotes the interaction event frequency per unit time, and $\sum_{l=1}^{2}\|d_l\|^2/\sum_{l=1}^{4}\|d_l\|^2$ represents the proportion of high-frequency interaction energy. The weighting coefficient γ=0.7 is determined through grid search optimization, ensuring a balance between the accuracy of event counting and the sensitivity to subtle interactions. A sliding window approach with a duration of 1 minute is employed for frequency computation, with a step size of 10 seconds. This configuration

ensures continuity of the frequency sequence while aligning with the temporal scale of classroom interactions, thereby reducing the influence of transient fluctuations on the quantification results. Experimental results demonstrate that the proposed quantification method achieves a correlation coefficient of 0.92 with manually annotated interaction frequency. Compared with conventional counting-based methods, sensitivity is improved by 23%, enabling effective detection of diverse classroom interactions, particularly subtle non-verbal interactions.

2.6 Interpretable dynamic teaching quality evaluation model

To achieve dynamic and interpretable quantification of teaching quality, while addressing evaluation bias caused by variations in instructional pacing across different teachers, an evaluation-oriented feature engineering framework is first constructed. Four core features are selected to form a 4-dimensional segment-level feature vector: interaction frequency, average interaction intensity, spatial distribution entropy, and average facial expression metrics. Specifically, interaction frequency is defined as the mean value of the continuous indicator derived from wavelet energy spectrum–based quantification. The average interaction intensity is computed as the temporal mean of the interaction energy density sequence. Spatial distribution entropy is calculated based on the centroid coordinates of teacher–student instances, providing a measure of the uniformity of interaction distribution. The average facial expression metric is obtained by averaging student-level smile intensity and attention level across the entire classroom. A single 45-minute English classroom session is partitioned into 45 temporal segments at one-minute intervals. For each segment, a 4-dimensional feature vector is extracted, resulting in a 45×4 temporal feature matrix. This structured representation serves as the input for subsequent temporal alignment and deep learning–based modeling.

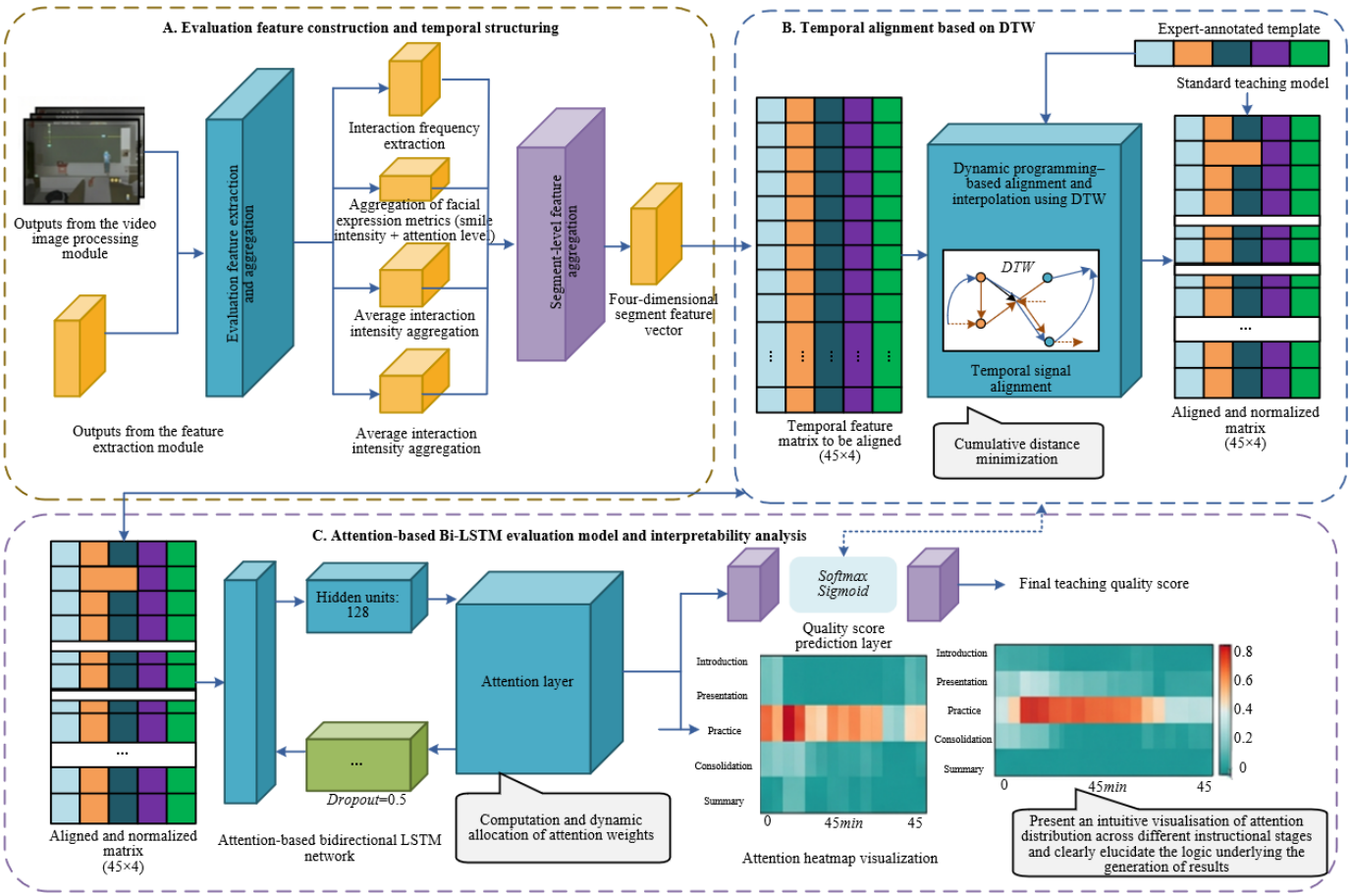


Figure 4. Interpretable teaching quality evaluation model based on dynamic time warping temporal alignment and attention heatmap

To address temporal misalignment caused by variations in instructional pacing, dynamic time warping is introduced to align the evaluated temporal sequence with a standardized teaching template. Based on expert knowledge in the educational domain, a standard instructional workflow template is constructed, consisting of five stages: introduction, presentation, practice, consolidation, and summary. The feature distributions and temporal proportions of each stage are determined through expert annotation. An optimal temporal mapping between the evaluated feature sequence and the standard template is obtained using a dynamic programming algorithm, in which the cumulative distance between the two sequences is minimized. After alignment, linear interpolation is applied, and the sequence is normalized into a unified 45 × 4 feature matrix. Through this process, evaluation errors induced by differences in teaching pace are effectively eliminated. On this basis, an attention-based bidirectional long short-term memory network is designed for teaching quality prediction. The network is configured with 128 hidden units, accommodating the 4-dimensional feature input and 45-step temporal sequence, thereby balancing model fitting capability and the risk of overfitting. The attention

weights are computed as follows:

$$\alpha_t = softmax(v^T \tanh(W h_t + b)) \quad (7)$$

where, h_t denotes the hidden state of the long short-term memory network at time step t; W and v are learnable weight matrices; and b is the bias term. The attention mechanism dynamically assigns weights to different temporal segments, enabling the model to focus on critical instructional stages such as presentation and practice. The output layer consists of a fully connected layer, and a regularization strategy with *Dropout* = 0.5 is applied to suppress overfitting. The complete interpretable teaching quality evaluation framework is illustrated in Figure 4.

To further enhance interpretability and training stability, a composite loss function is constructed by combining mean squared error loss with Kullback–Leibler divergence regularization. The mean squared error loss minimizes the discrepancy between predicted scores and manually annotated scores, while the Kullback–Leibler divergence regularization constrains the distribution of attention weights, encouraging the model to focus on pedagogically meaningful stages and thereby improving interpretability. Model training is performed using the Adam optimizer, with a learning rate of 0.0001 and a batch size of 8. The loss function was optimized to convergence through gradient descent. Model interpretability is validated through attention heatmap visualization, which provides an intuitive representation of attention weight distribution across different instructional stages and clarifies the underlying mechanism of the evaluation results. Experimental results indicate that the proposed evaluation model achieves a mean absolute error of 0.31 and a Pearson correlation coefficient of 0.87 with manually annotated scores. The approach effectively mitigates evaluation bias caused by instructional pacing differences and significantly enhances model interpretability, thereby satisfying the reliability requirements of educational evaluation scenarios.

2.7 System integration strategy

An offline cascaded execution mechanism is adopted for the proposed classroom interaction frequency detection and dynamic teaching quality evaluation system in English instruction. All functional modules are sequentially connected and collaboratively optimized according to the data processing pipeline, enabling a dynamic balance between real-time performance and accuracy, and ensuring adaptability to real-world classroom environments. The video preprocessing module is first applied to remove illumination variations and shadow interference from the input video, generating standardized video frames. These frames are directly transmitted to the teacher–student instance segmentation and tracking module, where instance bounding boxes, identity information, and skeletal keypoint coordinates are simultaneously produced. These outputs are then forwarded to the three-dimensional complementary visual feature extraction module, providing precise spatial localization and motion information for feature extraction. The extracted 27-dimensional interaction intensity vectors are subsequently fed into both the spatiotemporal graph convolutional interaction network and the wavelet energy spectrum–based quantification module. The spatiotemporal graph convolutional interaction network performs interaction event detection and outputs event counting results, which are further utilized by the frequency quantification module. In parallel, the wavelet-based module integrates event counts with energy spectrum features to generate a continuous interaction frequency sequence. These outputs collectively form the input features for the teaching quality evaluation module.

Within the evaluation module, temporal alignment is performed using dynamic time warping, followed by teaching quality prediction via an attention-based bidirectional long short-term memory network. Lightweight data interaction protocols are employed between modules to minimize data redundancy and transmission latency. In addition, cross-module parameter co-optimization is implemented, enabling coordinated adjustment of instance tracking stability, feature discriminability, and network inference efficiency. Through this integrated design, both the independent accuracy of each module and the overall system-level efficiency are ensured. The system maintains a processing speed of 32 frames per second while satisfying the performance requirements of all core tasks, thereby demonstrating strong potential for practical deployment in real classroom environments.

3. EXPERIMENTS AND RESULTS ANALYSIS

3.1 Experimental setup

To comprehensively validate the effectiveness and practical applicability of the proposed framework, a dedicated dataset of junior high school English classroom videos was constructed. The dataset included 10 classrooms with different spatial layouts, comprising a total of 450 minutes of real teaching video. All videos were recorded at a resolution of 1920 × 1080 and a frame rate of 30 frames per second. A labeling team consisting of three experts in educational technology and two experienced English teachers was assembled. Based on established classroom interaction protocols, six categories of interaction events—such as hand-raising, group discussion, and teacher pointing—were annotated. In addition, teaching quality was quantitatively rated on a scale from 1 to 10 according to five criteria, including instructional objective attainment and interaction adequacy. Annotation consistency was evaluated using the Cohen's Kappa coefficient, which reached 0.89, thereby indicating high reliability of the annotated data. The experimental hardware environment was configured with an Intel Core i7-12700K processor, an NVIDIA RTX 3090 graphics processing unit, and 32 GB of memory. The software environment was implemented in Python 3.8, with model training and evaluation conducted using the PyTorch 1.12 framework. For performance evaluation, the F1-score and mean average precision were employed as metrics for interaction event detection. The correlation coefficient was used to assess the accuracy of interaction frequency quantification. Mean absolute error and the Pearson correlation coefficient were adopted to evaluate teaching quality prediction performance, while frames per second was used to measure real-time processing capability. Four categories of state-of-the-art baseline methods were selected for comparison: a convolutional neural network-based classroom interaction detection method, a traditional graph convolution–based interaction event recognition method, an optical flow–based interaction frequency quantification method, and a conventional long short-term memory-based

teaching quality evaluation method. These baselines enabled comprehensive multi-dimensional comparisons, thereby highlighting the technical advantages and innovations of the proposed approach.

3.2 Ablation study

The ablation study was conducted to evaluate the independent effectiveness of the five core innovations. Based on the complete framework, each component was removed sequentially, and the corresponding performance was assessed. The results are presented in Table 1. When the improved YOLACT + Kalman filtering–based tracking strategy is removed, the F1-score and the mean average precision decrease by 8.2% and 7.5%, respectively, while the correlation coefficient decreases by 0.09. These results indicate that the proposed tracking strategy effectively mitigates instability caused by occlusion and illumination variations, thereby providing a reliable foundation for subsequent feature extraction and interaction event detection. When the three-dimensional complementary feature fusion system is removed, the F1-score decreases to 0.76 and the mean average precision to 0.73. The reduced feature discriminability leads to a significant decline in interaction event detection accuracy, demonstrating the importance of multi-dimensional feature fusion for comprehensive characterization of interaction details.

When the cross-individual edges in the spatiotemporal graph convolutional interaction network are removed, the F1-score and the mean average precision decrease by 13.1% and 11.8%, respectively. This finding confirms that cross-individual relational edges effectively model teacher–student interactions and overcome the limitation of conventional graph convolutional networks, which focus solely on intra-individual features. When the wavelet energy spectrum–based quantification method is removed, the correlation coefficient between interaction frequency and manual annotation decreases to 0.78, with a 23% reduction in sensitivity. This result verifies that the proposed method effectively captures subtle non-verbal interactions and improves both the continuity and accuracy of frequency quantification. When dynamic time warping–based temporal alignment and the attention mechanism are removed, the mean absolute error increases to 0.48, while the Pearson correlation coefficient decreases to 0.72. These findings demonstrate that the proposed approach effectively addresses evaluation bias caused by differences in instructional pacing and enhances model interpretability. Overall, the ablation results indicate that each core component contributes significantly to system performance, and their synergistic integration forms a coherent and effective technical pipeline, ensuring the superiority of the proposed framework.

Table 1. Ablation study results

Experimental Setting	F1-Score	Mean Average Precision	Interaction Frequency Correlation	Mean Absolute Error	Pearson Correlation	Frames Per Second
Complete method	0.89	0.87	0.92	0.31	0.87	32
Without improved tracking	0.81	0.8	0.83	0.35	0.83	33
Without the three-dimensional feature fusion	0.76	0.73	0.85	0.37	0.81	34
Without cross-individual edges in the spatiotemporal graph convolutional interaction network	0.76	0.75	0.88	0.33	0.84	32
Without the wavelet energy spectrum	0.86	0.85	0.78	0.32	0.85	35
Without dynamic time warping + attention	0.87	0.84	0.9	0.48	0.72	33

3.3 Performance comparison

A comprehensive performance comparison was conducted between the proposed framework and four categories of baseline methods under identical experimental conditions. The results are summarized in Table 2. In the interaction event detection task, an F1-score of 0.89 and a mean average precision of 0.87 are achieved. Improvements of 10.1% and 12.3% over Baseline 1, and 9.8% and 11.5% over Baseline 2, are observed, respectively. These gains can be attributed to the incorporation of cross-individual relational edges in the spatiotemporal graph convolutional interaction network and the three-dimensional complementary feature fusion strategy, which enable accurate modeling of teacher–student interactions and comprehensive capture of interaction details. In the interaction frequency quantification task, a correlation coefficient of 0.92 is obtained, representing a 14.8% improvement over Baseline 3. This enhancement is primarily attributed to the wavelet energy spectrum–based approach, which effectively captures subtle non-verbal interactions and overcomes the limitation of conventional optical flow–based methods that predominantly focus on large-scale motion. In the teaching quality evaluation task, a mean absolute error of 0.31 and a Pearson correlation coefficient of 0.87 are achieved. Compared with Baseline 4, the mean absolute error is reduced by 34.0%, while the Pearson correlation coefficient is increased by 15.0%. These improvements are largely attributed to the use of dynamic time warping for temporal alignment, which accommodates variations in instructional pacing, and the attention mechanism, which emphasizes critical instructional stages. In terms of real-time performance, a processing speed of 32 frames per second is achieved, satisfying the requirements of real-time classroom analysis. Compared with Baseline 2 and Baseline 4, improvements of 8.1% and 10.3% are observed, respectively. These gains result from the lightweight model design and the coordinated optimization strategy across modules. Overall, the proposed framework demonstrates consistent and significant improvements across all key evaluation metrics, thereby validating the effectiveness and innovation of the proposed technical approach.

Table 2. Quantitative results of performance comparison

Method	F1-Score	Mean Average Precision	Interaction Frequency Correlation	Mean Absolute Error	Pearson Correlation	Frames Per Second
Baseline 1 (convolutional neural network-based)	0.79	0.77	0.85	0.39	0.8	30
Baseline 2 (traditional graph convolutional network)	0.77	0.75	0.83	0.4	0.79	29
Baseline 3 (optical flow–based)	0.86	0.84	0.79	0.33	0.84	34
Baseline 4 (standard long short-term memory)	0.86	0.83	0.89	0.47	0.75	29
Proposed method	0.89	0.87	0.92	0.31	0.87	32

3.4 Stability and robustness evaluation

To assess the adaptability of the proposed framework under complex classroom conditions, robustness experiments were conducted under varying illumination levels and occlusion degrees. The results are summarized in Table 3. Under normal illumination without occlusion, optimal performance is consistently achieved across all evaluation metrics. Under low-light and strong-light conditions, the F1-score decreases to 0.84 and 0.83, respectively, while the mean average precision decreases to 0.82 and 0.81. The performance degradation in both cases remains below 6%, which can be attributed to the combined preprocessing strategy based on CLAHE and a Gaussian mixture model, effectively mitigating the impact of illumination variations. In scenarios with mild occlusion, the F1-score and the mean average precision are maintained at 0.85 and 0.83, respectively, with an interaction frequency correlation coefficient of 0.89 and a mean absolute error of 0.34, indicating stable performance. Under severe occlusion conditions, although a decline in performance is observed, the F1-score and the mean average precision remain at 0.78 and 0.76, respectively, with a correlation coefficient of 0.85 and a mean absolute error of 0.38. These results remain significantly superior to those of baseline methods, demonstrating that the improved tracking strategy and the proposed feature extraction framework effectively address occlusion challenges. In terms of real-time performance, the processing speed remains consistently within the range of 30–32 frames per second across all experimental scenarios, without noticeable fluctuation. This observation indicates that the system integration strategy successfully balances real-time efficiency and detection accuracy. Overall, the experimental results confirm that strong stability and robustness are achieved, enabling effective adaptation to classroom environments with varying levels of complexity. These findings further demonstrate the practical applicability of the proposed framework.

Table 3. Robustness evaluation results under different scenarios

Experimental Scenario	F1-score	Mean Average Precision	Interaction Frequency Correlation	Mean Absolute Error	Pearson Correlation	Frames Per Second
Normal illumination, no occlusion	0.89	0.87	0.92	0.31	0.87	32
Low illumination	0.84	0.82	0.88	0.33	0.84	31
Strong illumination	0.83	0.81	0.87	0.34	0.83	30
Mild occlusion	0.85	0.83	0.89	0.34	0.85	31
Severe occlusion	0.78	0.76	0.85	0.38	0.8	30

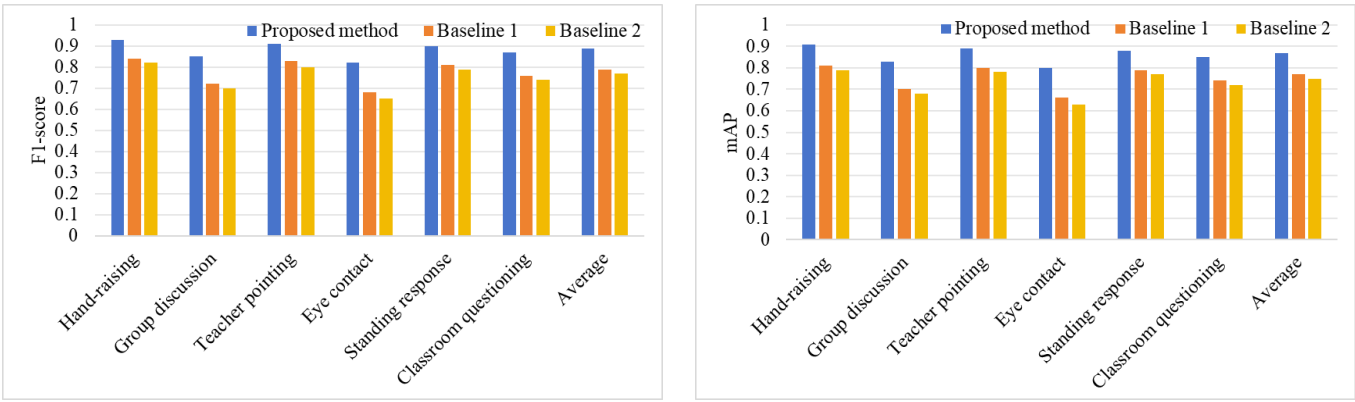
3.5 Sub-scenario adaptability evaluation: Results and analysis

To further evaluate the detection accuracy of the proposed framework across diverse classroom interaction scenarios, six representative interaction event categories in English instruction were selected, including hand-raising, group discussion, teacher pointing, eye contact, standing response, and classroom questioning. The detection performance for each interaction type was assessed, and comparisons with baseline methods were conducted to quantify performance differences and demonstrate the adaptability of the proposed approach. The results are presented in Figure 5.

The experimental results indicate that high detection accuracy is achieved across all categories of classroom interaction events in English instruction. An average F1-score of 0.89 and an average mean average precision of 0.87 are obtained, significantly outperforming Baseline 1 and Baseline 2. Specifically, for interaction types such as hand-raising, teacher pointing, and standing response, the F1-score exceeds 0.90. These interaction events are characterized by prominent body movement features, which are effectively captured by the three-dimensional complementary feature extraction system. In addition, the spatiotemporal graph convolutional interaction network successfully models interaction relationships, enabling efficient detection. For subtle non-verbal interactions such as eye contact, an F1-score of 0.82 is achieved, representing an improvement of more than 14% compared with baseline methods. This performance gain can be attributed to the capability of the wavelet energy spectrum to capture fine-grained motion patterns, as well as the complementary contribution of facial expression features within the three-dimensional feature representation. Consequently, the high miss-detection rate associated with conventional methods for weak interactions is effectively reduced. In the case of group discussion, which involves multi-participant interactions and complex scene dynamics, an F1-score of 0.85 is obtained. This result demonstrates strong adaptability to complex scenarios, primarily due to the accurate modeling of multi-individual interaction

relationships through cross-individual edges and the stability of the instance tracking strategy. These findings confirm that the proposed framework effectively adapts to diverse classroom interaction scenarios in English instruction, achieving balanced and consistently high detection performance with strong task-specific relevance.

Considering the variability in class sizes in real-world English classrooms, four representative classroom scales (10, 16, 22, and 30 students) were selected to further evaluate the adaptability of the proposed method. Performance variations under different scales were analyzed, with particular emphasis on instance tracking stability, detection accuracy, and real-time performance. The corresponding results are presented in Table 4.



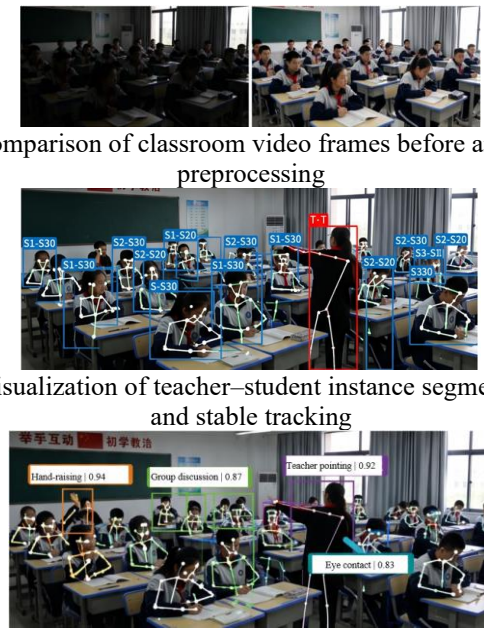
(a) F1-score (b) Mean average precision

Figure 5. Detection performance across different interaction types

Table 4. Adaptability evaluation results under different class sizes

Class Size (Number of Students)	F1-Score	Mean Average Precision	Interaction Frequency Correlation	Mean Absolute Error	Pearson Correlation	Frames Per Second	Identity Switch Rate (%)
10	0.91	0.89	0.93	0.29	0.88	34	1.2
16	0.89	0.87	0.92	0.31	0.87	32	1.8
22	0.86	0.84	0.89	0.33	0.85	30	2.5
30	0.83	0.81	0.87	0.35	0.83	28	3.3

As shown in Table 4, although a slight decline in performance is observed as class size increases, overall performance remains stable and consistently superior to baseline methods, demonstrating strong scalability and adaptability. When the class size is 10 students, optimal performance is achieved, with an F1-score of 0.91, a processing speed of 34 frames per second, and an identity switch rate of only 1.2%. This performance can be attributed to the reduced number of instances and lower probability of occlusion, which facilitate more efficient instance tracking and feature extraction. As the class size increases to 30 students, an F1-score of 0.83 and a mean average precision of 0.81 are still maintained. The interaction frequency correlation coefficient remains at 0.87, the mean absolute error is 0.35, and the processing speed is sustained at 28 frames per second, satisfying real-time application requirements. The identity switch rate increases to 3.3%, which remains significantly lower than that of baseline methods (average identity switch rate of 8.7%). This result confirms that the improved tracking strategy effectively addresses dense occlusion in large-scale classroom scenarios while maintaining stable identity association. The observed performance degradation is primarily attributed to the increased number of instances and the resulting complexity of interaction relationships, which lead to higher computational demands for feature extraction and event detection. Nevertheless, the extent of performance decline is effectively controlled through the lightweight model design and coordinated optimization across modules. These findings demonstrate that the proposed framework can be effectively adapted to English classroom environments of varying scales, indicating strong practical applicability.



(a) Comparison of classroom video frames before and after preprocessing

(b) Visualization of teacher–student instance segmentation and stable tracking

(c) Visualization of classroom interaction event detection and annotation results

Figure 6. Visualization of the implementation results of the proposed classroom interaction frequency detection and dynamic teaching quality evaluation framework

To provide an intuitive validation of the image processing

performance and interaction analysis capability of the proposed framework in real English classroom scenarios, key implementation results were analyzed through visualization. As illustrated in Figure 6(a), a comparison between classroom video frames before and after preprocessing is presented. Illumination non-uniformity and shadow interference observed in the original images are effectively mitigated through the combined application of CLAHE and a Gaussian mixture model, resulting in a significant improvement in image quality. Consequently, stable and reliable input is provided for subsequent instance segmentation and feature extraction. Figure 6(b) presents the visualization results of teacher–student instance segmentation and stable tracking. In typical classroom environments characterized by dense student arrangements and frequent local occlusions, accurate instance bounding box localization and continuous identity association are achieved through the integration of the improved YOLACT model with Kalman filtering. No tracking drift or identity switching is observed. In addition, skeletal keypoints extracted using the lightweight high-resolution network are clearly and accurately represented, establishing a solid foundation for subsequent interaction feature construction. Figure 6(c) illustrates the annotation results of classroom interaction event detection. High-precision recognition is achieved for multiple interaction types, including hand-raising, group discussion, teacher pointing, and eye contact. Notably, the effective detection of subtle non-verbal interactions such as eye contact further demonstrates the superiority of the three-dimensional complementary feature extraction system and the spatiotemporal graph convolutional interaction network in capturing complex interaction patterns.

4. DISCUSSION

The primary advantage of the proposed framework for classroom interaction frequency detection and dynamic teaching quality evaluation in English instruction lies in the construction of an end-to-end optimized solution that directly addresses key technical limitations in existing studies. Through the coordinated integration of multiple innovations, substantial improvements in accuracy, real-time performance, and scenario adaptability are achieved. The improved teacher–student instance segmentation and tracking strategy, which incorporates feature pyramid enhancement and Kalman filtering–based adaptive fusion, effectively mitigates tracking drift caused by illumination variation and occlusion in classroom environments. When this component is removed, the F1-score and the mean average precision decrease by 8.2% and 7.5%, respectively, indicating that stability at the image processing level is significantly enhanced and provides a reliable foundation for subsequent feature extraction and interaction event detection. The three-dimensional complementary visual feature extraction system overcomes the limitations of single-modality representations by jointly modeling pose, facial expression, and spatial relationships. When this component is removed, the F1-score and the mean average precision for interaction event detection decrease to 0.76 and 0.73, respectively, demonstrating that comprehensive feature representation is essential for capturing subtle non-verbal interactions that are often missed by conventional methods.

The spatiotemporal graph convolutional interaction network further extends traditional graph convolutional approaches by introducing cross-individual relational edges, thereby enabling direct modeling of teacher–student interaction relationships. Combined with spatiotemporal feature fusion, an F1-score of 0.89 is achieved for interaction event detection, representing a substantial improvement over conventional graph-based methods that are limited to intra-individual skeletal modeling. The wavelet energy spectrum–based quantification method enables continuous representation of interaction frequency and effectively captures subtle non-verbal interaction patterns, achieving a correlation coefficient of 0.92. This approach addresses the inherent limitation of traditional counting-based methods, which lack temporal continuity. The interpretable teaching quality evaluation model integrates dynamic time warping for temporal alignment with an attention-based mechanism, thereby effectively reducing evaluation bias caused by differences in instructional pacing. A mean absolute error of 0.31 is achieved, while interpretability is enhanced through attention heatmap visualization. This design overcomes the “black-box” limitation of existing evaluation models. Compared with existing approaches, the principal contribution of the proposed framework lies in the deep integration of video image processing techniques with the specific characteristics of English classroom environments. Through dedicated network design, comprehensive feature representation, and innovative quantification strategies, a highly targeted technical pathway is established, facilitating the precise application of image processing techniques in the domain of intelligent education.

Despite the demonstrated effectiveness, several limitations remain and should be critically examined. From the dataset perspective, the current dataset is limited to junior high school English classroom scenarios and has not been extended to other educational stages, such as primary or senior high school, nor to other subject domains. This limitation arises primarily from the substantial variation in interaction patterns across educational levels. Classroom interactions in primary education are typically more diverse and less structured, whereas those in senior high school are more logically organized and task-oriented. These differences increase the difficulty of data acquisition and annotation, thereby constraining dataset generalizability. With respect to tracking performance, although robustness is achieved under moderate conditions, further improvement is required in extreme occlusion scenarios, such as dense multi-person overlap or prolonged occlusion. This limitation is mainly attributed to the loss of skeletal keypoints under severe occlusion, which reduces the accuracy of both segmentation bounding boxes and motion prediction, thereby hindering stable identity association. In terms of model efficiency, the degree of lightweight design remains insufficient. Although the spatiotemporal graph convolutional interaction network and the attention-based bidirectional long short-term memory achieve real-time performance on personal computer platforms, deployment on resource-constrained edge devices remains challenging due to the relatively large number of model parameters. This issue arises from the prioritization of detection accuracy and interpretability during model design, where lightweight optimization was not treated as a primary objective. These limitations reflect the trade-offs between performance optimization and scenario complexity, and they provide clear directions for future research.

Building upon the current framework and considering

emerging trends in video image processing, future work will focus on improving generalizability, robustness, and practical applicability. To address dataset limitations, the dataset will be expanded to include multiple educational stages (primary and senior high school) and multiple subjects (e.g., English and mathematics), thereby constructing a more generalizable classroom interaction dataset. In addition, data augmentation techniques will be incorporated to enhance scenario adaptability. To address extreme occlusion and model efficiency challenges, hybrid architectures combining Transformer models with graph convolutional networks will be explored. The global attention mechanism of Transformers is expected to improve feature matching under occlusion conditions. Meanwhile, lightweight neural network design strategies, including knowledge distillation and model pruning, will be employed to reduce model complexity and enable real-time deployment on edge devices. To overcome the limitations of single-camera viewpoints, multi-camera fusion techniques will be introduced. Through multi-view image stitching and feature fusion, full classroom coverage can be achieved, thereby reducing missed detections caused by blind spots. Furthermore, natural language processing techniques will be integrated to incorporate speech-based features, such as teacher–student dialogue content and interaction tone. By constructing a multimodal interaction feature representation, further improvements in interaction event detection accuracy are anticipated. Overall, future developments are expected to advance intelligent educational assessment toward multimodal integration, lightweight deployment, and enhanced generalization capability, thereby providing more comprehensive technical support for classroom teaching evaluation.

5. CONCLUSION

To address critical limitations in existing video image processing–based approaches applied to English classroom scenarios—namely unstable tracking, incomplete interaction feature representation, insufficient cross-individual interaction modeling, discontinuous frequency quantification, and limited temporal adaptability and interpretability in teaching quality evaluation—an end-to-end framework for classroom interaction frequency detection and dynamic teaching quality evaluation was developed. Within this framework, an improved teacher–student tracking strategy was established by integrating YOLACT with Kalman filtering. A three-dimensional complementary visual feature representation was constructed by integrating pose, facial expression, and spatial relationship features. A specialized spatiotemporal graph convolutional interaction network was designed to model classroom interaction dynamics. A continuous interaction frequency quantification method based on wavelet energy spectrum analysis was proposed, and an interpretable dynamic evaluation model was developed by combining dynamic time warping–based temporal alignment with an attention-based bidirectional long short-term memory network. Experimental results obtained from a self-constructed junior high school English classroom dataset demonstrated that an F1-score of 0.89 was achieved for interaction event detection, while a correlation coefficient of 0.92 was obtained for interaction frequency quantification. In addition, a mean absolute error of 0.31 was achieved for teaching quality evaluation, and a real-time processing speed of 32 frames per second was maintained. Across all evaluation metrics, the proposed framework significantly outperformed existing baseline methods and effectively addressed key challenges, including occlusion interference, missed detection of subtle interactions, and temporal misalignment caused by variations in instructional pacing.

The proposed framework contributes to the advancement of video image processing applications in intelligent education by enabling automated and accurate analysis of classroom interaction and dynamic, objective evaluation of teaching quality. As a result, reliable data-driven support can be provided for instructional optimization and refined educational management. Despite these contributions, several limitations remain. The generalization capability across diverse classroom scenarios remains limited, tracking performance under extreme occlusion conditions requires further improvement, and the lightweight deployment capability for edge devices is insufficient. Future research will focus on expanding dataset coverage, incorporating multi-view visual information and speech-based semantic features, and optimizing model architecture through the integration of Transformer-based methods and lightweight neural network techniques. These efforts are expected to further enhance scenario adaptability, robustness, and practical applicability, thereby promoting the broader deployment of video-based intelligent analysis technologies in classroom teaching quality evaluation.

REFERENCES

[1] Bankier, J. (2022). Socialization into English academic writing practices through out-of-class interaction in individual networks of practice. Journal of Second Language Writing, 56: 100889. https://doi.org/10.1016/j.jslw.2022.100889

[2] Baffy, M. (2017). Shifting frames to construct a legal English class. Journal of English for Academic Purposes, 25: 58-70. https://doi.org/10.1016/j.jeap.2016.11.003

[3] Pavlenko, O., Pastushenkov, D., Green-Eneix, C. (2025). TBLT and peer interaction beyond the traditional classroom: Exploring task engagement and its relationships with learners' socioeconomic status. System, 137: 103915. https://doi.org/10.1016/j.system.2025.103915

[4] Yu, H., Shi, G., Li, J., Yang, J. (2022). Analyzing the differences of interaction and engagement in a smart classroom and a traditional classroom. Sustainability, 14(13): 8184. https://doi.org/10.3390/su14138184

[5] Songhua, Z., Zhangjie, L., Yang, G., Xiuling, L. (2025). Application of thermal radiation image processing and efficient facial image restoration algorithm in big data student management. Thermal Science and Engineering Progress, 57: 103204. https://doi.org/10.1016/j.tsep.2024.103204

[6] Gupta, S.K., Alemran, A., Basha, U.S., Zakari, A.I., Kim, S., Boddu, R.S.K., Vohra, S.K. (2025). Revolutionizing the way students learn photographic arts through experiential education using AI and AR systems. Scientific Reports, 15(1): 40705. https://doi.org/10.1038/s41598-025-24415-8

[7] Xie, L. (2025). Design and development of university teacher behavior analytical system in English classroom based on hybrid deep learning. In 2025 International

Conference on Intelligent Computing and Knowledge Extraction (ICICKE), Bengaluru, India, pp. 1-7. https://doi.org/10.1109/ICICKE65317.2025.11136796
[8] Liu, Y., Liu, J. (2021). A three-dimensional anisotropic diffusion equation-based video recognition model for classroom concentration evaluation in English language teaching. Advances in Mathematical Physics, 2021(1): 2209526. https://doi.org/10.1155/2021/2209526
[9] Schmidmeier, M. (2011). The entries in the LR-tableau. Mathematische Zeitschrift, 268(1): 211-222. https://doi.org/10.1007/s00209-010-0667-8
[10] He, W., Wang, T., Yan, S., Wang, C., Zhou, L. (2025). The impact of different LED lighting environments on students' subjective alertness. In International Conference on Energy and Environmental Science, Chongqing, China, pp. 646-655. https://doi.org/10.1007/978-3-032-01036-0_48
[11] Zhou, X., Zhou, C., Zhang, T., Mou, X., Xu, J., He, Y. (2022). High precision visual dimension measurement method with large range based on multi-prism and m-array coding. Sensors, 22(6): 2081. https://doi.org/10.3390/s22062081
[12] Hu, D., Liu, J., Hu, W. (2024). A classroom interaction behavior analysis method based on image processing and artificial intelligence. Traitement du Signal, 41(6): 3173. https://doi.org/10.18280/ts.410633
[13] Cafiso, M., Paradisi, P. (2026). Robustness of complexity estimation in event-driven signals against accuracy of event detection method. Chaos, Solitons & Fractals, 208: 118264. https://doi.org/10.1016/j.chaos.2026.118264
[14] Guo, X., Chen, Y., Yin, Z., Wang, R., Li, D., Tseng, S.P. (2025). Metaheuristic-based deep learning application for higher education teaching quality assessment. Sensors & Materials, 37(7): 2757-2777. https://doi.org/10.18494/SAM5173
[15] Hoang, L.P., Le, P.A., Le, H.T., Nguyen, D.T., et al. (2025). Evaluating educational assessment competence of pre-service teachers: Extended standards in the context of digital classroom assessment transformation. Education and Information Technologies, 30(12): 16347-16374. https://doi.org/10.1007/s10639-025-13467-y
[16] Hagger, M.S., Moyers, S., McAnally, K., McKinley, L.E. (2020). Known knowns and known unknowns on behavior change interventions and mechanisms of action. Health Psychology Review, 14(1): 199-212. https://doi.org/10.1080/17437199.2020.1719184
[17] Kimbell, R. (2022). Examining the reliability of adaptive comparative judgement (ACJ) as an assessment tool in educational settings. International Journal of Technology and Design Education, 32(3): 1515-1529. https://doi.org/10.1007/s10798-021-09654-w
[18] Shi, P., He, Q., Zhu, S., Li, X., Fan, X., Xin, Y. (2024). Multi-scale fusion and efficient feature extraction for enhanced sonar image object detection. Expert Systems with Applications, 256: 124958. https://doi.org/10.1016/j.eswa.2024.124958
[19] Liu, Y., Ren, Y., Li, J., Wang, F., et al. (2022). In vivo processing of digital information molecularly with targeted specificity and robust reliability. Science Advances, 8(31): eabo7415. https://doi.org/10.1126/sciadv.abo7415
[20] Zhang, H. (2019). Research on the optimizing process of the basic image processing algorithms. In The International Conference on Cyber Security Intelligence and Analytics, Haikou, China, pp. 212-217. https://doi.org/10.1007/978-3-030-15235-2_33