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

Traitement du Signal
Vol. 42, No. 3, June, 2025, pp. 1383-1391
Journal homepage: http://iieta.org/journals/ts

Computer Vision-Based Analysis of Teacher–Student Interaction Behaviors in Dynamic English Classroom Environments

Yuhui Wang

Basic Teaching Department, Henan Logistics Vocational College, Zhengzhou 453500, China

Corresponding Author Email: 13213025957@163.com

https://doi.org/10.18280/ts.420314

Received: 17 December 2024
Revised: 9 May 2025
Accepted: 5 June 2025
Available online: 30 June 2025

Keywords:
computer vision, dynamic classroom environment, English language instruction, teacher–student interaction behavior, multiview fusion

ABSTRACT

Under the framework of educational informatization, dynamic classroom environments—characterized by flexible spatial layouts and integrated multimedia interaction technologies—have emerged as pivotal settings for English language instruction. Teacher–student interaction, as a core component of the instructional process, encompasses diverse verbal and non-verbal communication patterns that significantly influence both pedagogical effectiveness and student learning experiences. The advancement of computer vision enables more precise analysis of such interactive behaviors; however, current approaches remain limited. Methods relying on single-view visual data often fail to comprehensively capture the complexity of classroom interactions, resulting in incomplete behavioral recognition. Furthermore, traditional handcrafted feature extraction techniques have shown limited capacity in capturing dynamic motion cues, thereby reducing recognition accuracy in diverse classroom postures and gestures. To address these limitations, a multiview fusion-based method for recognizing teacher–student interaction behaviors in English classrooms was proposed. The model comprises a multiview fusion module and a motion feature extraction module. The former integrates visual information from multiple camera angles to mitigate single-view occlusion and information loss, while the latter leverages deep learning to effectively model dynamic bodily expressions and postural transitions. Experimental results demonstrate that the proposed method significantly enhances recognition accuracy of interactive behaviors in dynamic classroom environments. This approach provides a robust data-driven foundation for optimizing instructional strategies, supporting personalized learning, and informing intelligent classroom design. The findings contribute both theoretical insights and practical value to the integration of computer vision technologies in English language education.

1. INTRODUCTION

With the advancement of educational informatization, dynamic classroom environments—characterized by flexible spatial configurations, abundant multimedia teaching resources, and real-time interactive educational equipment [1-4]—have increasingly been adopted as mainstream settings for English language instruction. Within these environments, teacher–student interaction behaviors exhibit diverse and dynamic patterns [5, 6], encompassing not only traditional verbal exchanges but also non-verbal cues such as gestures, facial expressions, and body language. These interaction behaviors constitute a fundamental component of English teaching processes and exert a substantial influence on both instructional effectiveness and student learning experience. Concurrently, the rapid development of computer vision technologies [7-10] provides robust technical support for the accurate capture and analysis of teacher–student interactions in dynamic classroom settings, enabling in-depth exploration of such behaviors from a visual perspective.

The analysis of teacher–student interaction behaviors in dynamic English classroom environments holds significant practical implications. From the instructional perspective, the accurate recognition and interpretation of these behaviors can offer teachers real-time, visual feedback, facilitating timely adjustments to teaching strategies and pedagogical methods. Such adjustments can optimize instructional processes and enhance overall classroom quality. For instance, by examining the relationship between teachers' gestural and verbal explanations and students' responsive behaviors [11, 12], a more accurate assessment of students' engagement levels and learning needs can be obtained, thus improving the precision and effectiveness of instruction. From the student learning perspective, a deeper understanding of interaction behaviors can reveal learning patterns and characteristics of students during classroom interactions. This, in turn, can provide a basis for personalized learning support and learning effectiveness evaluation, ultimately contributing to the enhancement of students' overall English language proficiency. Furthermore, insights derived from such analyses can inform the design and optimization of dynamic classroom environments, thereby fostering the integration of educational technology with English language instruction.

Although significant progress has been made in the analysis

of teacher–student interaction behaviors within classroom environments, several limitations remain. Some studies [13-16] have primarily relied on single-view visual information, which has proven insufficient for capturing the complexity of interactions occurring in dynamic classroom environments. This reliance has led to partial and sometimes fragmented recognition and analysis of interaction behaviors. For example, a single camera view may fail to simultaneously capture both instructional activities performed by the teacher at the lectern and responsive behaviors exhibited by students in their seating areas, resulting in the omission of critical interactive details. In terms of motion feature extraction, certain studies [17-20] have adopted traditional handcrafted feature extraction techniques, which demonstrate limited effectiveness in identifying dynamic movement patterns during teacher–student interactions. It is difficult for these methods to accommodate the diversity of human postures and movements in dynamic classroom settings, thereby constraining the accuracy and robustness of interaction behavior analysis.

A method for recognizing teacher–student interaction behaviors in English classroom environments based on multiview fusion was proposed in this study. The constructed model consists of two primary modules: a multiview fusion module and a motion feature extraction module. The multiview fusion module integrates visual information captured from multiple camera perspectives to obtain a more comprehensive view of teacher–student interactions, thereby overcoming the limitations of single-view systems. The motion feature extraction module employs advanced deep learning techniques to efficiently extract and model dynamic motion characteristics observed during interactive behavior, enhancing the model's capacity to recognize complex interaction patterns with higher accuracy. The value of this research lies in the integration of multiview fusion and efficient motion feature extraction, which enables more accurate identification of teacher–student interaction behaviors in dynamic English classroom environments. This enhanced recognition capability provides high-resolution behavioral data to support the in-depth analysis of interaction patterns and the optimization of instructional processes. Moreover, the proposed method is expected to offer new methodological perspectives and technical references for related studies, advancing the field of computer vision–based educational behavior analysis. The approach also holds broad application prospects in practical English language instruction.

2. MULTIVIEW FUSION FOR RECOGNIZING TEACHER–STUDENT INTERACTIONS IN ENGLISH CLASSROOMS

As a representative complex interactive setting, the dynamic classroom environment comprises multiple functional zones, including the lecture area, student seating area, and group discussion area. Teacher–student interaction behaviors within such environments are often characterized by cross-regional and multimodal features. Teachers may turn to face students while writing on the board or engage in close-range communication during in-class supervision, whereas students may participate through actions such as raising hands, asking questions, or engaging in collaborative group work. A single-camera perspective can capture only a partial view of the classroom scene, frequently resulting in the loss of critical visual cues—such as students' micro-expressions and group-level gestural coordination—thus limiting the ability to reconstruct the dynamic interaction chain of "teacher instruction–student feedback–bidirectional regulation" typical of English classroom discourse. However, in a computer vision–based multiview acquisition strategy, cameras are deployed at various locations, such as the classroom ceiling and side walls, to synchronously capture multiple visual dimensions, including frontal views of the teacher's instructional posture, lateral views of student reactions, and panoramic activity trajectories. This setup enables comprehensive 3D visual coverage of teacher–student interaction behaviors. The resulting heterogeneous and multisource visual data serve as the essential foundation for accurately identifying complex behavioral patterns in English classrooms, such as question–answer exchanges, group discussions, and emotional engagement. Multiview fusion techniques function as the critical technological bridge for transforming fragmented visual inputs into coherent behavioral representations.

2.1 Model architecture design

A method for recognizing teacher–student interaction behaviors in English classroom environments based on multiview fusion was proposed. The overall model architecture is illustrated in Figure 1.

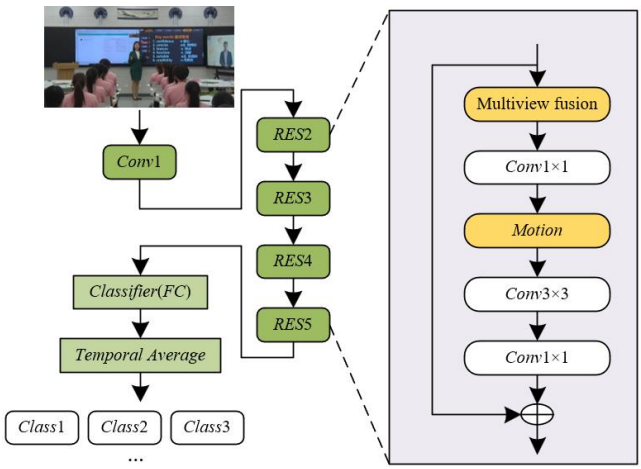

Figure 1. Overall architecture of the multiview fusion-based recognition model for teacher–student interaction behaviors in English classroom environments

In dynamic classroom environments, a multiview computer vision data acquisition setup was implemented by deploying panoramic ceiling-mounted cameras, close-up side-view lenses at the lectern, and wide-angle cameras on the classroom walls. These devices were used to synchronously capture multiview video streams that comprehensively cover teacher–student interaction scenarios. To ensure temporal and spatial coherence, spatiotemporal alignment techniques were applied during preprocessing to synchronize the multisource visual data across different camera perspectives. Following alignment, the multiview fusion module performs cross-view modeling using a feature-level fusion strategy. This process involves integrating 2D pose features—such as OpenPose keypoint coordinates—with appearance features (e.g., clothing color, teaching aids), as well as scene-context features

(e.g., blackboard content, seating layout). For example, during a teaching aid demonstration, fine-grained hand movements are captured by the lectern-side camera, while the panoramic view records collective student gaze directions and head postures. The fusion module employs an attention mechanism to adaptively weight salient features from each view, thereby generating a comprehensive visual representation that encodes spatial positioning, coordinated body movements, and semantic scene context. This approach effectively resolves the issue of interaction detail loss caused by blind spots in single-view visual data.

After the completion of spatial multiview feature fusion, a multi-scale temporal attention mechanism was introduced to model the long-range temporal dynamics of video segments. This mechanism is specifically designed to capture the sequential nature of interaction behaviors in English classrooms, such as the staged action chain observed during question–answer segments (e.g., "teacher poses a question – student raises hand – student stands to respond"). To implement this, continuous video streams were segmented into clips of varying temporal resolutions, including short (e.g., 1-second), medium (e.g., 10-second), and long (e.g., full-lesson) segments. A bidirectional temporal difference network was then employed to compute motion differences across segments, enabling the identification of key frames associated with significant interaction events, such as sudden student movements (e.g., standing up) or abrupt changes in teacher gestures. Subsequently, an attention mechanism-based temporal weighting map was generated, allowing the model to focus selectively on critical interaction intervals—such as turn-taking moments in dialogue or instances of emotional resonance—while suppressing the influence of irrelevant routine actions, such as page-turning or blackboard cleaning. In group discussion scenarios, for example, the mechanism captures continuous interaction sequences such as "teacher approaches discussion group – leans in to listen – nods in response." These temporally correlated multiview fusion features were enhanced through attention-based weight reinforcement and were ultimately integrated with segment-level motion features, resulting in a dynamic motion representation that encodes long-term temporal dependencies.

2.2 Multiview fusion module

Due to the spatiotemporal complexity of teacher–student interaction behaviors in dynamic English classroom environments—such as the motion trajectories of teachers delivering instruction while walking and the postural shifts of students engaged in multidirectional interactions—traditional video analysis methods that rely solely on height–width planar modeling have proven inadequate. These conventional approaches often fail to capture the deeper coupling between temporal (S) and spatial dimensions. Dynamic behaviors, including horizontal teacher movement from the lectern to student groups and vertical actions such as students standing to respond, are frequently segmented into isolated spatial frame sequences when analyzed on a single plane. As a result, temporal dependency information tends to be lost. To address this limitation, the proposed multiview fusion module introduces two novel analytical dimensions: height–time (G×S) and width–time (Q×S). Together with the traditional spatial plane, these dimensions constitute a 3D modeling framework. Within this framework, triaxial spatiotemporal convolutional kernels—3×1×1, 1×3×1, and 1×1×3—are employed to capture dynamic features along the temporal, horizontal, and vertical axes, respectively. This design overcomes the limitations of conventional methods in mining temporal dynamics, establishing a technical pathway for accurately decoding the spatiotemporal correlations among actions, motion trajectories, and classroom scenes within English instruction. The architecture of the multiview fusion module is illustrated in Figure 2.

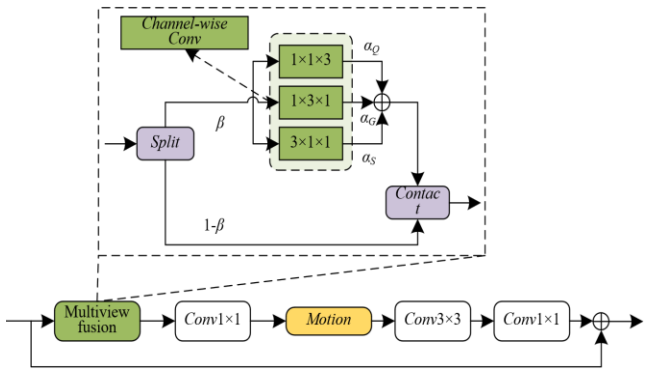

Figure 2. Architecture of the multiview fusion module

The module is designed based on an input feature tensor $A \in R^{Z \times S \times G \times Q}$, which is initially divided along the channel dimension into two branches: the raw activation branch $A^1 \in R^{\beta Z \times S \times G \times Q}$, and the multiview modeling branch $A^2 \in R^{(1-\beta)Z \times S \times G \times Q}$. The parameter β controls the proportion between raw information retention and spatiotemporal modeling information, thereby optimizing the trade-off between feature completeness and computational efficiency. For A^2, multiview modeling is achieved through three directional channel convolutions:

Temporal convolution (*kernel size*: 3×1×1): This operation slides along the temporal axis S, capturing sequential frame-to-frame changes in behaviors such as the evolution of teacher gestures or the dynamic shifts in student facial expressions. This enables the modeling of temporal dependencies in staged interaction patterns such as "questioning–thinking–responding."

Horizontal convolution (*kernel size*: 1×3×1): This operation focuses on variations along the width axis Q, facilitating the analysis of horizontal trajectories such as teacher walking paths or lateral student group movements during collaborative tasks.

Vertical convolution (*kernel size*: 1×1×3): This operation targets motion along the height axis G, capturing key posture changes, including student transitions between sitting and standing, or vertical body movements during teacher blackboard writing.

The outputs of the three directional convolutions—denoted as P_S, P_G, and P_Q—were fused via weighted summation across dimensions to generate a high-order representation that integrates multiview spatiotemporal information. This representation was subsequently concatenated with the raw activation branch A^2, resulting in a composite feature representation that preserves both local detail and global dynamic structure. Let z, s, a, and b represent the indices along the channel, temporal, height, and width dimensions, respectively. The convolution kernels used for modeling the G-Q, Q-S, and G-S views are denoted by J^S, J^G, and J^Q. The expressions for the outputs of the three directional convolutions are given by:

$$P_S = \sum_u J_{z,u}^S \Pi A_{z,s+u,g,q}^1 \quad (1)$$

$$P_G = \sum_u J_{z,u}^G \Pi A_{z,s,g+u,q}^1 \quad (2)$$

$$P_Q = \sum_u J_{z,u}^Q \Pi A_{z,s,h,q+u}^1 \quad (3)$$

Let σ denote the activation function, and let the activated feature map be represented as $P^1 \in R^{\beta Z \times S \times G \times Q}$. The corresponding weights for each view are denoted as α_S, α_G, and α_Q, respectively. The fusion of P_S, P_G, and P_Q was performed using the following equation:

$$P^1 = \sigma\left(\alpha_S \cdot P_S + \alpha_G \cdot P_G + \alpha_Q \cdot P_Q\right) \quad (4)$$

In this study, α_S=α_G=α_Q=1 was set. Finally, the outputs were concatenated along the channel dimension—denoted as *concat*—to produce the output of the multiview fusion module:

$$D_u = concat(A^2, P^1) \quad (5)$$

where, $D_u \in R^{V \times Z \times S \times G \times Q}$. In the context of English classroom interaction analysis, the triaxial modeling capabilities of the proposed module yield substantial advantages. For instance, when a teacher engages in a compound behavior involving both teaching aid demonstration and blackboard explanation in the lectern area, the width–time convolution captures the temporal correlation between horizontal movement and blackboard content updates. Simultaneously, the height–time convolution identifies vertical motion trajectories such as the raising or lowering of instructional tools, while the original spatial-plane convolution retains spatial relations, including fine-grained details of the teaching aid and the collective gaze direction of the students. The fusion of these three components enables precise semantic parsing of multistage interaction events, such as "teacher points to blackboard vocabulary → students engage in collective repetition → teacher nods in affirmation." In group discussion scenarios, the horizontal convolution can be used to track the teacher's movement between groups, the vertical convolution detects height changes as students stand to speak, and the temporal convolution captures gesture exchanges and text-passing sequences, along with their temporal intervals. Together, these fused features construct a comprehensive feature vector that encodes spatial positioning, action sequencing, and interaction rhythm.

2.3 Motion feature extraction

Given the temporal dynamics inherent in teacher–student interaction behaviors within dynamic English classroom environments—such as multi-stage action chains in classroom questioning or sustained collaborative patterns during group discussions—traditional video analysis methods that rely on direct computation of long-range frame differences are prone to introducing noise and often overlook the temporal dependencies within intermediate steps. This can result in distorted representations of motion patterns associated with complex interaction behaviors. To address these limitations, the motion feature extraction module was designed around a "neighboring segment temporal-difference modeling" strategy. A multi-scale receptive field (MF) mechanism was introduced to capture fine-grained motion information and to mitigate discontinuities in long-term motion representation. The architecture of the motion feature extraction module is illustrated in Figure 3. Specifically, the input to this module consisted of the fused feature map sequence D_u, obtained from the multiview fusion module. The continuous video was first segmented into equal-length clips. Motion differentials between adjacent segments—such as frame-wise optical flow fields and pose keypoint displacements—were then computed to generate a short-term motion difference matrix. This approach avoids the computational complexity and redundancy associated with long-range frame comparisons, while precisely focusing on the dynamic progression of interaction sequences such as "question – response – feedback," thereby providing a temporal characteristic basis for the frequent occurrence of verbal exchanges accompanied by coordinated physical gestures in English classrooms.

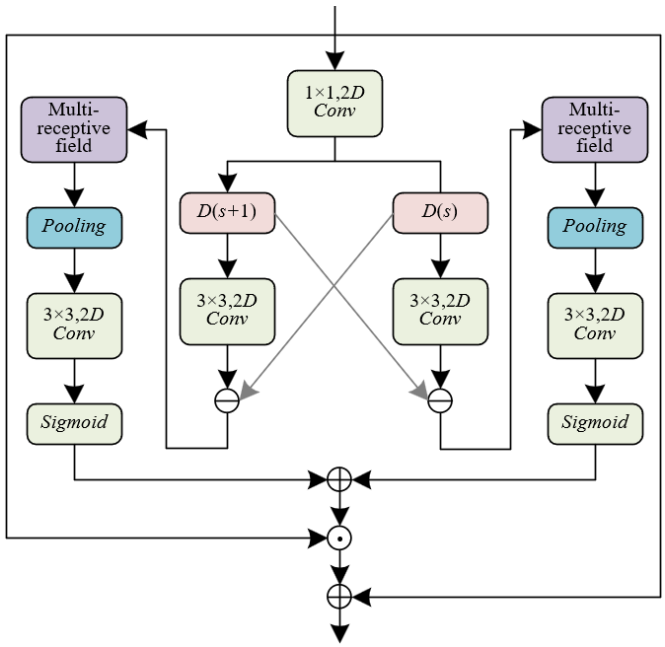

Figure 3. Architecture of the motion feature extraction module

A multi-receptive field strategy based on a three-branch parallel architecture was adopted within the module to perform multi-scale feature extraction on the motion difference matrix. This was achieved using dilated convolutions with varying dilation rates, enabling multi-level capture of fine-grained motion information under a shared-weights and lightweight design.

Small receptive field branch (*dilation rate* = 1): This branch focuses on capturing instantaneous frame-to-frame motion changes, such as subtle finger tremors during teaching aid manipulation or student micro-expressions like blinking. These features help preserve local interaction behavior details.

Medium receptive field branch (*dilation rate* = 2): Designed to cover a moderate temporal span of approximately 5–10 frames, this branch is suitable for analyzing the temporal correlations of staged action sequences, such as "teacher walks toward student → pauses for explanation" or "student raises hand → stands to answer."

Large receptive field branch (*dilation rate* = 3): This branch processes long temporal sequences extending beyond 20

frames, focusing on the interaction rhythms of an entire lesson. Examples include the initiation–peak–resolution cycle of group discussions and fluctuations in instructional tempo over the course.

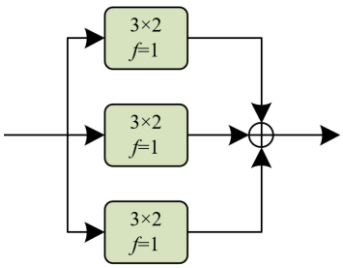

Figure 4. Architecture of the multi-receptive field module

The architecture of the multi-receptive field module is shown in Figure 4. Motion features extracted by each branch were temporally aligned through zero-padding and then concatenated to form a composite motion feature vector. This vector integrates instantaneous motion details, mid-range stage transitions, and long-range interaction patterns. Subsequently, an attention mechanism was introduced to generate an attention map over the long-term motion features. This mechanism selectively amplifies features corresponding to key interactive moments—such as physical contact between teacher and student or eye-gaze alignment—while suppressing irrelevant motions such as equipment shifts or student posture adjustments. In this way, a transformation is achieved from raw motion signals to semantically enriched features.

Formally, the temporal alignment difference between video segments D^e_u is denoted by $T(D^e_u, D^e_{u+1})$, and is defined as:

$$T(D^e_u, D^e_{u+1}) = J_2 * D^e_{u+1} - D^e_u \tag{6}$$

The preliminary long-term motion information extracted from video segment u via the multi-receptive field module was then used to compute the attention map, denoted by $L(D^e_u, D^e_{u+1})$, where:

$$L(D^e_u, D^e_{u+1}) = \frac{1}{3}\sum_{k=1}^{3} Fk\left(\bar{T}(D^e_u, D^e_{u+1})\right) \tag{7}$$

Finally, the attention map for long-term motion features within each segment was computed via residual connection, expressed as:

$$K(D^e_u, D^e_{u+1}) = SIG\left(J_3\left(MAX\left(L(D^e_u, D^e_{u+1})\right)\right)\right) \tag{8}$$

$$L(D^e_u, D^e_{u+1}) = \frac{1}{2}\left[K(D^e_u, D^e_{u+1}) + K(D^e_{u+1}, D^e_u)\right] \tag{9}$$

In complex interactive scenes of English classroom environments, the residual connection architecture is employed to effectively mitigate gradient vanishing issues encountered in the training of deep networks, thereby ensuring the stability of multi-stage feature extraction. For example, in analyzing the compound action sequence of "blackboard writing followed by turning to pose a question," the small receptive field branch captures fine wrist movements during writing; the medium receptive field branch models torso posture changes during the turning action; and the large receptive field branch integrates the entire temporal span from the beginning of the writing to the end of the questioning behavior. Through residual connection, these features are complementarily fused, avoiding fragmentation caused by single-scale modeling. In group discussion scenarios involving simultaneous participation by multiple students, the attention map dynamically focuses on the standing actions and gestural interactions of active speakers, while also retaining the spatial trajectory of the teacher's supervisory movement. This yields a multidimensional motion representation encompassing spatial positioning, temporal sequencing, and interaction roles.

3. EXPERIMENTAL RESULTS AND ANALYSIS

As shown in Table 1, the proposed method achieved a Top-1 accuracy of 81.6%, significantly outperforming all comparative approaches. Specifically, models such as Temporal Difference Networks (TDN) (73.6%) and Temporal Shift Module (TSM) (74.6%)—which rely on single-view inputs or basic temporal modeling—exhibited limitations in capturing spatial dynamics across multi-zone classroom interactions, leading to suboptimal performance. In contrast, the multiview fusion module adopted in the present approach employed cross-dimensional convolutions over front, side, and rear perspectives, thereby enabling comprehensive spatial distribution modeling of teacher–student interactions. This led to an improvement of 7–8 percentage points over single-view methods, confirming the necessity of multiview information integration. Although methods such as SlowFast Networks (72.8%) and Expanded 3D Convolution Network (X3D) (75.9%) utilize 3D convolutions or dual-path structures, their capacity for fine-grained motion feature extraction remains limited. By contrast, the motion feature extraction module of the proposed model—through temporal difference computation between adjacent segments and multi-scale dilated convolution—effectively distinguished multi-level motion patterns such as gesture emphasis by teachers and cyclical group discussions. This enabled superior modeling of the coupled "language–gesture–expression" behaviors typical of English classroom interactions, resulting in a 5.7 percentage point improvement over X3D. These results validate the synergistic advantage of combining temporal difference modeling with multi-receptive fields. Furthermore, models such as Global Spatial-Temporal Attention (GSTA) (72.8%) and Non-local Network (72.4%) demonstrated insufficient scene-specific adaptability in their attention mechanisms, making it difficult to focus on key interaction events such as teacher–student physical contact or manipulation of instructional tools. In contrast, the joint view–motion attention design employed in the proposed method successfully suppressed non-informative scenes while enhancing the semantic salience of pedagogically meaningful interactions. This led to an 8.8 percentage point improvement over GSTA, highlighting the model's high degree of contextual compatibility with educational settings.

As shown in Table 2, the performance differences observed before and after the introduction of the multiview fusion module clearly demonstrate its critical contribution to model enhancement. Specifically, under short temporal window conditions, the Top-1 accuracy increased from 72.56% to 72.89%, representing an absolute gain of +1.26%. This improvement indicates that during rapid interaction cycles—

such as teacher questioning followed by immediate student responses—the multiview fusion mechanism successfully integrated spatial cues from the front, side, and rear perspectives. This integration compensated for the blind spots inherent in single-view setups, enhanced the spatial contextual understanding of brief interactions, and improved recognition accuracy. In scenarios involving medium-to-long temporal durations, Top-1 accuracy improved from 72.56% to 74.58%, corresponding to a +1.38% gain. Under these conditions, the module not only fused spatial information but also modeled temporal relationships across views, thereby strengthening the joint spatiotemporal representation of extended interaction sequences. The synergistic effect between deeper backbone networks and the multiview fusion module was more pronounced, indicating that the multidimensional spatiotemporal features generated by the module were effectively exploited by deeper architectures. This facilitated accurate adaptation to the semantic complexity of classroom interactions. The ablation results in Table 2 confirm that the multiview fusion module effectively addresses the core challenge of information deficiency in single-view modeling under dynamic classroom conditions. Significant performance improvements were observed across both short-term and medium-to-long-term interaction scenarios. The module demonstrated strong adaptability to the spatial distribution and multidirectional characteristics of English classroom interactions and provided essential support for subsequent motion feature extraction. Its technical innovation and contextual alignment establish the multiview fusion mechanism as a central component in educational computer vision systems, offering the foundation for achieving high-performance interaction behavior recognition.

Table 1. Comparison between the proposed method and state-of-the-art approaches

Method	Pretrain	Backbone	*Top*-1
TDN	*Kinetics*-400+*Classroom*-100	*ResNet*-50	73.6%
TSM	*ImageNet*+*Kinetics*-600+*Classroom*-200	*ResNet*-50/101	74.6%
SlowFast Networks	*Kinetics*-700+*Classroom*-300	*SlowPath*:*ResNet*-50 *FastPath*:*ResNet*-101	72.8%
Non-local Network	*Kinetics*-400+*Classroom*-150	*ResNet*-50/101	72.4%
X3D	*Kinetics*-600+*Classroom*-200	*Lightweight* 3*DResNet*	75.9%
TPN	*Kinetics*-500+*Classroom*-250	*ResNet*-50	74.5%
GSTA	*YouTube*-8*M*+*Classroom*-100	*ResNet*-50	72.8%
Proposed method	*ImageNet*+*Kinetics*	*ResNet*-50	81.6%

Table 2. Ablation study: comparison of model performance before and after introducing the multiview fusion module

Method	Frames	Backbone	*Top*-1	*Top*-5	*ΔTop*-1
Before multiview fusion	9	*Resnet*-50	72.56	97.58	+1.26
After multiview fusion			72.89	97.23	
Before multiview fusion	14	*Resnet*-101	72.56	97.88	+1.38
After multiview fusion			74.58	98.36	

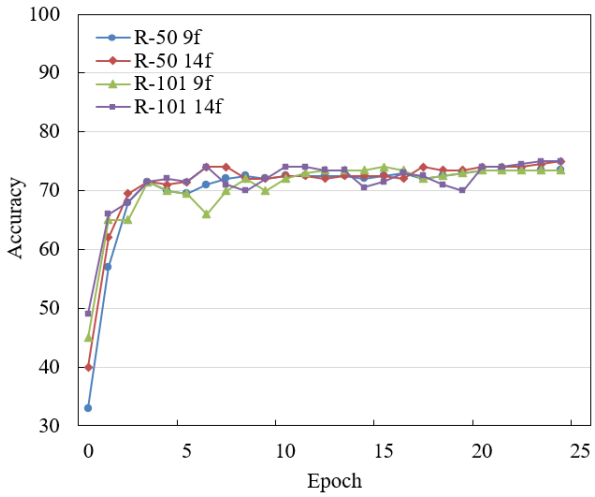

Figure 5. Recognition results of teacher–student interaction behaviors in English classroom environments under different network depths and frame inputs

Figure 5 illustrates the model convergence curves and final accuracy values under varying backbone network depths and input frame numbers. It can be observed that ResNet-101 consistently achieves higher accuracy than ResNet-50 in the later stages of training. This indicates that deeper networks are more capable of fully leveraging the cross-view spatiotemporal features generated by the multiview fusion module, thereby enabling more effective modeling of hierarchical interactive semantics in English classrooms, such as instruction–feedback–collaboration. For example, the ResNet-101 + 14-frame configuration surpassed 75% accuracy by the 25th training epoch, while the ResNet-50 + 9-frame configuration plateaued at approximately 70%, confirming the superior representational capacity of deeper networks in capturing the complex interaction patterns present in educational settings. Additionally, models with 14-frame input demonstrated faster convergence and higher final accuracy compared to those with 9-frame input. A longer temporal window provided the motion feature extraction module with richer dynamic cues, thereby enhancing its ability to model long-duration interaction patterns. For instance, in the task of capturing a teacher's movement trajectory across multiple student groups, the 14-frame configuration covered more stages of the interaction process, effectively reducing the risk of missing critical actions often encountered with shorter sequences and significantly improving recognition accuracy. The convergence analysis in Figure 5 confirms that the integration of deeper backbone networks with longer temporal inputs fully activates the spatial capabilities of the multiview fusion mechanism and the temporal capacity of the motion feature extraction. As a result, the proposed method exhibits efficient feature learning and semantic modeling in recognizing teacher–student interaction behaviors under

dynamic English classroom conditions. The combined use of deeper networks and extended temporal windows yielded notable performance improvements, further validating the educational specificity of the technical design. This strategy successfully addresses the challenges of limited single-view information and loss of dynamics in short sequences. By enabling hierarchical feature extraction, the proposed framework achieves precise recognition of complex interactions and provides a robust pathway for computer vision–based educational behavior analysis. The experimental evidence strongly supports the method's effectiveness.

Table 3. Ablation study: performance impact of the multi-receptive field mechanism

Model Configuration	Training Accuracy (%)	Testing Accuracy (%)
Without the multi-receptive field mechanism	95.6	75.8
Full model	96.2	77.2

As shown in Table 3, the multi-receptive field mechanism plays a critical role in enhancing model performance. On the training set, the full model achieved an accuracy of 96.2%, representing a 0.6 percentage point improvement over the configuration in which the mechanism was removed (95.6%). This improvement reflects the effectiveness of the parallel multi-scale convolution strategy in capturing detailed motion features from the training data. For example, when learning compound behaviors such as "teacher writing on the board + student note-taking," the small receptive field branch focuses on fine motor movements of the teacher's hand, the medium receptive field models the temporal sequence of note-taking behaviors, and the large receptive field integrates the overall instructional rhythm. This multi-scale capture enhances the model's ability to fit training data and prevents the loss of features often associated with single-scale modeling. The improvement was even more pronounced on the testing set, where the full model reached 77.2%, surpassing the 75.8% accuracy of the reduced model by 1.4 percentage points. This gain can be attributed to the diverse temporal scales present in interaction behaviors within dynamic classroom environments. Teacher–student interactions often include both instantaneous micro-movements and prolonged behavioral patterns. The multi-receptive field mechanism, through the use of dilated convolutions, enables the simulation of receptive fields over varying time spans, supporting the joint modeling of micro-level actions and macro-level temporal patterns. For instance, in the task of recognizing the action "student stands to answer," the small receptive field captures the abrupt change in posture at the moment of standing, while the large receptive field models the surrounding classroom context before and after the response. This joint encoding enhances the model's generalization capability, particularly in the presence of real-world challenges such as lighting variation and occlusion. The ablation results presented in Table 3 provide strong evidence that the multi-receptive field mechanism, through effective multi-scale feature extraction, significantly improves the recognition of teacher–student interaction behaviors in dynamic English classroom environments. The observed performance gains in both training and testing scenarios demonstrate the mechanism's ability to model multi-scale interactive semantics with high precision. Its deep integration within the overall model architecture addresses the limitations of single-scale feature representation and provides critical support for the recognition of complex interaction behaviors.

Table 4. Impact of input frame count and network depth on model performance

Backbone	Frames	Training Accuracy (%)	Testing Accuracy (%)
Resnet-50	9	75.6	74.5
	14	76.2	75.2
Resnet-101	9	75.8	75.9
	14	76.4	76.5

As presented in Table 4, a consistent upward trend in both training and testing accuracy was observed with increases in network depth and input frame count. For the testing set, ResNet-50 with 9-frame input achieved an accuracy of 74.5%, whereas ResNet-101 with 9-frame input reached 75.9%, demonstrating that deeper networks possess stronger abstraction capabilities for the multiview fusion features, thereby enabling improved capture of spatiotemporal dependencies in classroom teacher–student interactions.

Further, the configuration using ResNet-50 with 14-frame input yielded 75.2%, while ResNet-101 with 14-frame input reached 76.5%. These results indicate that longer temporal sequences provide the motion feature extraction module with more comprehensive dynamic information, thereby enhancing the modeling of long-duration interaction patterns and mitigating the loss of critical behavioral stages that often occurs in short-sequence inputs. The experimental findings in Table 4 confirm that the proposed method—through the joint design of deep backbone networks and long temporal sequence input—fully leverages the spatial capacity of the multiview fusion and the temporal ability of the motion feature extraction. This configuration enabled the accurate recognition of teacher–student interaction behaviors under dynamic English classroom conditions. The combined optimization of network depth and input frame count effectively addressed two core challenges: insufficient information in single-view inputs and loss of dynamics in short sequences. Through hierarchical feature extraction, a unified spatial–temporal–semantic representation was constructed. These results provide strong empirical support for the effectiveness of the proposed framework. The design offers a reliable technical pathway for computer vision–based analysis in educational environments, and its proven scene adaptability and performance advantages lay a solid foundation for future research.

4. CONCLUSION

To address the technical challenges of recognizing teacher–student interaction behaviors in dynamic English classroom environments, an end-to-end method integrating multiview spatial modeling with fine-grained temporal dynamic analysis was proposed. Two core modules were developed to support this framework. By deploying multiview cameras to capture classroom video data, channel separation and triaxial spatiotemporal convolution were employed to achieve cross-view feature fusion, thereby resolving the issue of interaction detail loss caused by single-view blind spots. To address the temporal dynamics of English classroom interactions, a multi-receptive field strategy based on temporal difference computation between adjacent segments was designed. A

three-branch parallel structure was constructed using dilated convolutions, enabling the model to capture instantaneous micro-movements, staged action sequences, and long-term interaction patterns. An attention mechanism was further integrated to enhance the weight of features associated with key interactive moments while suppressing noise from irrelevant scenes. The use of residual connections ensured stable training of the deep network and enabled efficient transformation from raw motion signals to semantically enriched features.

This research is tightly aligned with the multidirectional interaction characteristics of English classroom environments. Through multiview fusion, full spatial coverage of the classroom was achieved, allowing for the first time the spatiotemporal modeling of cross-perspective teacher mobility and multidirectional student feedback. The motion feature extraction module, combining adjacent segment difference calculation with multi-receptive fields, effectively addressed issues inherent in conventional methods such as high noise in long-range frame computations and disrupted temporal dependencies. Using hierarchical feature extraction across small, medium, and large receptive fields, the model was capable of capturing instantaneous features such as subtle changes in student mouth movement during repetition, as well as modeling long-term patterns such as the overall rhythm of a lesson. This provided a robust hierarchical feature representation of frequent language–gesture synchronized behaviors in educational scenarios. Ablation experiments confirmed that this mechanism improved test accuracy by 1.4% to 2.2%. These findings offer critical technical support for intelligent classroom analysis systems and enable automated semantic interpretation of teacher–student interactions, laying a foundation for applications such as instructional evaluation and structured retrieval of classroom video content. Compared to generic video analysis models, the proposed approach demonstrated superior performance in context-specific interaction recognition within educational environments, underscoring its substantial practical value.

However, several limitations in this study remain to be addressed. First, the current model relies on synchronized acquisition from multiple cameras, which introduces considerable complexity in hardware deployment and demands high spatial–temporal alignment accuracy. As a result, full adaptation to low-cost, single-camera classroom environments has not yet been achieved. Second, in scenarios involving severe occlusion, complex lighting conditions, or unstructured interactions, the model's ability to focus on critical features may be weakened, resulting in fluctuations in recognition accuracy. Third, the model training still depends on large-scale, manually annotated datasets specific to educational contexts, limiting generalization performance under low-resource or small-sample conditions. Future efforts may focus on developing lightweight multiview fusion approaches based on self-supervised learning, thereby reducing dependence on hardware. Enhancements in spatial information modeling under monocular settings should also be pursued to improve deployment flexibility. The integration of non-visual modalities such as audio and text is encouraged to construct cross-modal interaction models that couple linguistic and visual features, thereby enabling more comprehensive interpretation of compound behaviors such as spoken instructions accompanied by gestural demonstrations. Additionally, domain adaptation techniques could be explored to enhance model generalization across diverse classroom environments and instructional models.

ACKNOWLEDGEMENT

This paper was supported by General Project of Humanities and Social Sciences Program for Universities in Henan Province (Grant No.: 2025-ZDJH-282).

REFERENCES

[1] Fraschini, N. (2023). Language learners' emotional dynamics: Insights from a Q methodology intensive single-case study. Language, Culture and Curriculum, 36(2): 222-239. https://doi.org/10.1080/07908318.2022.2133137

[2] Elmimouni, H., Šabanović, S., Rode, J.A. (2024). Navigating the cyborg classroom: Telepresence robots, accessibility challenges, and inclusivity in the classroom. ACM Transactions on Accessible Computing, 17(2): 8. https://doi.org/10.1145/3672569

[3] Connor, C.M., Spencer, M., Day, S.L., Giuliani, S., Ingebrand, S.W., McLean, L., Morrison, F.J. (2014). Capturing the complexity: Content, type, and amount of instruction and quality of the classroom learning environment synergistically predict third graders' vocabulary and reading comprehension outcomes. Journal of Educational Psychology, 106(3): 762-778. https://doi.org/10.1037/a0035921.

[4] Kareema, M.I.F., Hakmal, M.H.M. (2025). ESL and ASL students' and teachers' perspectives on online classroom management techniques. Ijaz Arabi Journal of Arabic Learning, 8(1): 54-74. https://doi.org/10.18860/ijazarabi. V8i1.28867

[5] Doyle, N.B., Downer, J.T., Brown, J.L., Lowenstein, A.E. (2022). Understanding high quality teacher-student interactions in high needs elementary schools: An exploration of teacher, student, and relational contributors. School Mental Health, 14(4): 997-1010. https://doi.org/10.1007/s12310-022-09519-0

[6] Song, W., Zhang, C., Gao, M. (2022). Analysis method for teacher-student interaction in online English courses. International Journal of Emerging Technologies in Learning, 17(9): 170-183. https://doi.org/10.3991/ijet.v17i09.31371

[7] Mettes, P., Ghadimi Atigh, M., Keller-Ressel, M., Gu, J., Yeung, S. (2024). Hyperbolic deep learning in computer vision: A survey. International Journal of Computer Vision, 132(9): 3484-3508. https://doi.org/10.1007/s11263-024-02043-5

[8] Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.M. (2017). Computer vision for assistive technologies. Computer Vision and Image Understanding, 154: 1-15. https://doi.org/10.1016/j.cviu.2016.09.001

[9] Heidari, S., Dinneen, M.J., Delmas, P. (2024). Quantum annealing for computer vision minimization problems. Future Generation Computer Systems, 160: 54-64. https://doi.org/10.1016/j.future.2024.05.037

[10] Parvaiz, A., Khalid, M.A., Zafar, R., Ameer, H., Ali, M., Fraz, M.M. (2023). Vision transformers in medical computer vision—A contemplative retrospection. Engineering Applications of Artificial Intelligence, 122: 106126. https://doi.org/10.1016/j.engappai.2023.106126

[11] Ayvazo, S., Aljadeff-Abergel, E. (2014). Classwide peer tutoring for elementary and high school students at risk: Listening to students' voices. Support for Learning, 29(1): 76-92. https://doi.org/10.1111/1467-9604.12047
[12] Diaz, P., Hrastinski, S., Norström, P. (2024). How teacher students used digital response systems during student teaching. Education and Information Technologies, 30(7): 8953-8978. https://doi.org/10.1007/s10639-024-13165-1
[13] Akkaya, N. (2014). Elementary teachers' views on the creative writing process: An evaluation. Educational Sciences: Theory and Practice, 14(4): 1499-1504. https://doi.org/10.12738/estp.2014.4.1722
[14] Korur, F., Eryilmaz, A. (2018). Interaction between students' motivation and physics teachers' characteristics: Multiple case study. Qualitative Report, 23(12): 3054-3083.
[15] Yilmaz, K., Altinkurt, Y. (2009). Prospective teachers' views about occupational unethical behaviours. Turkish Journal of Business Ethics, 2(2): 71-88.
[16] Liu, B., Peng, B., Zhang, Z., Huang, Q., Ling, N., Lei, J. (2023). Unsupervised single-view synthesis network via style guidance and prior distillation. IEEE Transactions on Circuits and Systems for Video Technology, 34(3): 1604-1614. https://doi.org/10.1109/TCSVT.2023.3294521
[17] Tisnés, H.M. (2023). Main recent study topics on teacher-student interaction. Interdisciplinaria, 40(2): 23-40. https://doi.org/10.16888/interd.2023.40.2.2
[18] Pennings, H.J., van Tartwijk, J., Wubbels, T., Claessens, L.C., van der Want, A.C., Brekelmans, M. (2014). Real-time teacher-student interactions: A dynamic systems approach. Teaching and Teacher Education, 37: 183-193. https://doi.org/10.1016/j.tate.2013.07.016
[19] Sason, H., Kellerman, A. (2021). Teacher-student interaction in distance learning in emergency situations. Journal of Information Technology Education: Research, 20: 479-501. https://doi.org/10.28945/4884
[20] Pielmeier, M., Huber, S., Seidel, T. (2018). Is teacher judgment accuracy of students' characteristics beneficial for verbal teacher-student interactions in classroom? Teaching and Teacher Education, 76: 255-266. https://doi.org/10.1016/j.tate.2018.01.002