ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

An Intelligent Evaluation Model for English Classroom Interaction Quality Based on Multimodal Data

Yang Liu[1,*] *and Tao Yang*[2]

[1] Jilin Vocational College of Industry and Technology, Jilin, Jilin, 132000, China
[2] Zunyi Normal College, Zunyi, Guizhou, 563000, China

SUMMARY: *In response to the problem of strong subjectivity and insufficient process evidence in the evaluation of interactive quality in English classrooms, this paper constructs a multimodal intelligent evaluation model that integrates video, audio, and classroom transcribed text. Based on 48 real English classes and 1920 interactive segments, establish a five dimensional annotation system for questioning quality, feedback effectiveness, participation breadth, emotional atmosphere, and target language interaction density. The results showed that the intra class correlation coefficient (ICC) was 0.86. The macro average F1 value (Macro-F1) of the proposed model on the test set is 0.803, the mean absolute error (MAE) is 0.298, and the correlation coefficient between classroom level prediction and expert rating is 0.861. The ablation and case analysis show that there are differences in the dependence of different interaction dimensions on text, audio, and visual modalities. The model can also identify clues such as open questioning, waiting time, and participation coverage, providing interpretable basis for teachers to improve classroom interaction.*

KEYWORDS: *multimodal learning analytics; English classroom interaction; teaching quality evaluation; cross-modal fusion; interpretable artificial intelligence*

1 Introduction

The quality of English classroom interaction is not a marginal indicator in classroom evaluation, but a key link connecting teacher questioning, feedback organization, student participation, and target language practice. For a long time, this object has mainly relied on manual class evaluation, classroom observation scales, and post class questionnaires for judgment. This type of approach can provide some teaching experience support, but it is also generally constrained by the subjective judgment of evaluators, limited observation periods, and insufficient process evidence, making it particularly difficult to form stable and traceable evaluation results in large sample, long-term, and cross class contexts. For English classrooms that emphasize immediate response, interactive feedback, and language co construction, relying solely on post class impression based grading is no longer sufficient to fully reveal the true quality of interaction. Liu et al. constructed an intelligent evaluation model based on English teachers' classroom language interaction and emotional behavior, indicating that the automated recognition of classroom interaction has gradually moved from concept verification to feasible technical solutions, and also demonstrating that English classroom evaluation is shifting from "result judgment" to "process drawing" [1]. At the same time, the popularization of smart classrooms, classroom recording systems, and learning platforms

*esther9090@163.com
https://doi.org/10.65102/is2026809
Published: 30 April, 2026

enables classroom scenes to continuously generate multi-source data such as videos, audios, automatically transcribed text, and behavior logs. Compared with traditional observation methods, these data not only retain the chronological order of interactions, but also record fine-grained clues such as speaking rounds, facial expressions, attention allocation, and speech density, providing a more solid evidence basis for quantitative modeling of classroom interaction quality. Wang et al. used small object detection technology to evaluate the quality of classroom teaching in universities, demonstrating that visual computing methods can extract behavioral information related to teaching status in complex classroom backgrounds. This provides practical feasibility for capturing participation and interactive activity from classroom videos [2].

In broader research on teaching evaluation, the evaluation framework has also shifted from single score output to multi-source information integration. Zhou et al. proposed a multi domain heterogeneous data fusion evaluation framework for hybrid classrooms, emphasizing that teaching quality judgment cannot rely solely on a certain type of data, but should establish correlations between multimodal, multi scenario, and multi subject information [3]. This progress provides important insights for the study of the quality of English classroom interaction: truly effective evaluation should not be limited to general judgments about whether the classroom is good or not, but should further answer more specific questions such as "how teachers ask questions, how students respond, whether feedback promotes continuous interaction, and whether interaction is balanced among different students". For English teaching research, classroom interaction itself has distinct multimodal attributes. The waiting time after teachers ask questions, the timing for students to continue speaking, the direction of gaze and posture, and the organization of speech stress and pauses all jointly affect the way interaction progresses. Wang and Dai investigated multimodal teacher-student interactions in EFL classrooms through lagged sequence analysis, revealing identifiable sequential structures between verbal and nonverbal behaviors. This suggests that English classroom interactions do not occur randomly, but have dynamic patterns that can be modeled [4]. Therefore, constructing "English classroom interaction quality" as a separate operable evaluation object not only has theoretical necessity, but also has a methodological foundation. Furthermore, the quality of classroom interaction cannot be narrowed down to whether the language form is accurate. Querol Juli á n conducted a multimodal analysis of online teaching in English teaching environments and found that promoting student participation is not only about what teachers say, but also about how they use discourse organization, gaze, action, and technical interfaces to construct accessible interactive spaces. This means that high-quality English classroom interaction should simultaneously reflect the openness of questioning, the extensibility of feedback, the coverage of participation opportunities, and the supportive emotional atmosphere, all of which are difficult to fully reflect by a single textual feature. Recent research also suggests that the value of multimodal interaction analysis lies not only in "seeing the classroom clearly", but also in "improving the classroom". Zhou et al. pointed out through a case study of novice EFL teachers that multimodal interaction analysis can in turn influence teachers' teaching practices and identity construction, enabling teachers to have a clearer understanding of their own behavioral patterns in questioning, responding, and organizing participation. This indicates that if the intelligent evaluation model only outputs a high or low score, its educational significance is still limited; Only by identifying which interactive features drove high scores and which behavioral patterns caused low scores, can the evaluation results truly be transformed into feedback that teachers can use.

From the perspective of interactive micro mechanisms, the limitations of the unimodal perspective are becoming increasingly apparent. Badem's research on the initiation of video mediated EFL classroom correction suggests that teachers often rely on both screen resources

and body movements to guide self correction when dealing with students' hesitation, incorrect answers, or unsatisfactory responses. In other words, the "promotion" and "repair" in classroom interaction do not only occur at the oral level, but are embedded in the collaborative operation of speech, action, and interface clues. If the evaluation model relies solely on transcribed text, it may overlook the most critical evidence at the turning point of interaction. Similarly, the role of teacher gestures and body prompts in English classrooms should not be underestimated. Ş im ş ek Tontu ş and Kuru G ö nen found that in synchronous online language classrooms, teachers use embodied guidance strategies to engage students in speaking and maintaining participation. This discovery further indicates that visual information is not an ancillary variable, but a component of interaction quality assessment. From this, it can be seen that a truly intelligent evaluation model for the quality of English classroom interaction needs to establish complementary representations between visual, auditory, and textual aspects, rather than simply concatenating multiple features and directly providing classification results.

Despite the rapid development of multimodal learning analysis enhanced by artificial intelligence, relevant reviews have clearly pointed out several shortcomings that still exist in this field. The systematic review by Mohammadi et al. indicates that existing research generally faces problems such as insufficient real-life classroom scenarios, rough modal fusion, weak interpretability, and limited implementation of educational applications [9]. Corresponding to the evaluation of the quality of English classroom interaction, there are at least four gaps that have not been fully addressed: firstly, many studies focus on the overall evaluation of classroom teaching quality, but have not broken down the quality of English classroom interaction into operational and interpretable evaluation dimensions; Secondly, existing models often remain at the level of single modal analysis or simple feature concatenation, lacking detailed modeling of the complementary relationship between vision, speech, and text; Thirdly, existing work has placed more emphasis on indicators such as accuracy and precision, but has provided less explanation on what interactive evidence the model relies on to make judgments, and there has been less discussion on how the results can serve teacher improvement; Fourthly, many studies have not adequately controlled for teacher or class overlap in data partitioning, resulting in models potentially learning individual teacher styles rather than more universal interaction patterns, leading to biased results. At the application level, multimodal learning analysis has also begun to enter teacher decision support systems. Possaghi et al. introduced multimodal learning analysis dashboards into K-12 education, demonstrating that multi-source classroom data can assist teachers in classroom scheduling and real-time decision-making [10]. However, such systems focus more on classroom monitoring and visualization presentation, and for the specific object of "English classroom interaction quality", there is still a lack of an integrated evaluation path from fragment level recognition to classroom level grading, and then to the interpretation of key influencing factors. That is to say, existing research has demonstrated the usefulness of multimodal data, but has not yet fully answered the question of how to transform this data into a stable evaluation model for English classroom interaction quality.

Based on this, this article intends to construct a multimodal intelligent evaluation model for English classroom interaction quality, which takes classroom videos, voice signals, and automatically transcribed text as inputs, outputs interaction quality scores at both segment and classroom levels, and further identifies key factors that affect the evaluation results. Unlike treating classroom evaluation as a single label, this article defines the quality of English classroom interaction as a composite construct that is decomposable, interpretable, and feedbacked, focusing on core dimensions such as questioning quality, feedback effectiveness, participation breadth, emotional atmosphere, and target language interaction density. The

main contributions of this article are reflected in three aspects. Firstly, propose a five dimensional evaluation framework for the quality of English classroom interaction, transforming the originally vague classroom impression evaluation into observable, annotatable, and computable analysis objects. Secondly, construct a visual audio text cross modal fusion model to characterize the complementarity of different modalities in online expression, and improve the accuracy and robustness of evaluation in complex classroom contexts. Thirdly, verify the effectiveness of the model from four aspects: accuracy, robustness, interpretability, and teaching feedback value, striving to make the evaluation results not only "score", but also provide teachers with actionable classroom improvement basis

2 Methods

2.1 Multimodal English classroom dataset construction and annotation

This study focuses on constructing a multimodal classroom dataset for the quality of English classroom interaction. The original lesson examples were collected from September 2024 to June 2025, sourced from real English classrooms in three universities and one middle school. There were a total of 48 lessons, involving 12 teachers, with each teacher providing 4 complete lesson examples. The course types cover comprehensive English, English reading, English speaking, and listening and speaking training, with a single effective teaching time of about 40 minutes and a total duration of about 32 hours. In order to preserve the original appearance of classroom interaction as much as possible, the collection end synchronously records four types of information: panoramic videos obtained from fixed machine positions at the back of the classroom, close-up videos for teachers, environmental audio and teacher clip on microphone audio, as well as classroom transcription text generated by an automatic speech recognition system and manually verified. All videos are uniformly processed to 1080p, 25 fps, and audio is uniformly converted to 16 kHz mono to ensure consistency in cross modal alignment in the future. The multimodal English classroom interaction quality evaluation method adopted in this article covers data collection, preprocessing, segment segmentation, quality annotation, feature extraction, cross modal fusion, and classroom level evaluation output. The overall framework is shown in Figure 1.

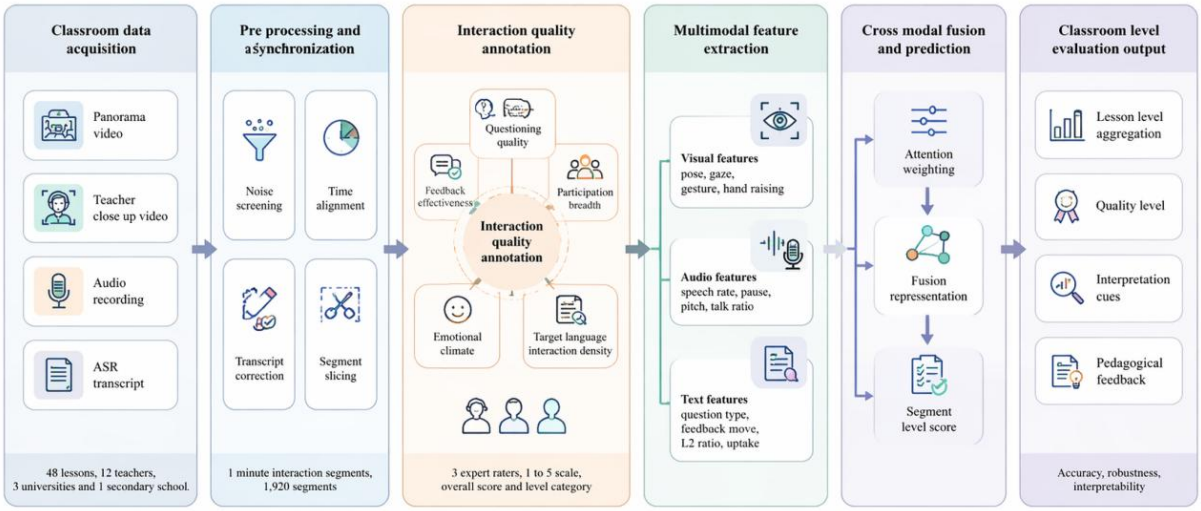


Figure 1: Overall framework of multimodal classroom interaction quality evaluation.

As shown in Figure 1, the multimodal English classroom interaction quality evaluation method adopted in this article first collects panoramic classroom videos, teacher close-up videos, audio signals, and transcribed text. Subsequently, data cleaning, time alignment, and fragment segmentation were completed, and based on this, a five dimensional interactive quality label was constructed [11]. Then, further extract visual, audio, and text features, and generate segment level scores through cross modal fusion models. Finally, summarize the results of the segments at the classroom level to form a comprehensive evaluation of the quality of English classroom interaction. In the preprocessing stage, the research team first removed segments with equipment abnormalities, prolonged occlusion, severe noise, or a high proportion of non teaching activities, and then performed continuous slicing in units of 1 minute, obtaining a total of 1920 interaction segments. When slicing, preserve the original timestamp and synchronize the alignment of panoramic video, teacher close-up, dual channel audio, and text transcription [12]. For segments with overlapping speeches between teachers and students, mixed English Chinese expressions, board instructions, and a large number of oral fillers, the "ASR initial transcription+manual review" method is used for correction to reduce errors in text feature extraction.

This study breaks down English classroom interaction quality into five interpretable dimensions. Specifically, questioning quality focuses on whether the questioning has openness, progressiveness, and depth of inquiry. Feedback effectiveness examines whether teacher responses can form clarification, expansion, or reorganization. Participation breadth measures whether speaking opportunities are distributed among a wider group of students, rather than being concentrated among a few active students. Emotional climate describes the emotional atmosphere of encouragement, acceptance, tension, or indifference during the interactive process. Target language interaction density measures the proportion, continuity, and functional load of English as an interactive medium. Each dimension is scored using a 1-5 scale, and the five dimensional mean is calculated as the total score of the segment, which is further mapped to 5 levels for dual validation in subsequent classification and regression experiments [13].

The annotation work was completed by three experts with backgrounds in English teaching methods or applied linguistics. Before formal annotation, 120 fragments are randomly selected for trial scoring, and the scoring manual is repeatedly revised based on divergent examples until the scoring criteria are basically stable. When grading formally, the annotator first browses the entire lesson in its entirety, and then returns to the segment level for detailed grading. To avoid misjudgment caused by isolated fragments, each rating allows for a 15 second review of the context before and after the current fragment. If the difference in a certain dimension exceeds 1 point, enter the negotiation round and record the reason for revision. The final results showed that the intra group correlation coefficient ICC of the five dimensional total score was 0.86, and the weighted Kappa of each dimension ranged from 0.72 to 0.81, indicating that the annotation system has good stability and repeatability [14, 15]. In order to facilitate the centralized presentation of data sources, collection modes, fragment composition, and five dimensional annotation definitions, this article summarizes the dataset composition and label design in Table 1.

Table 1: Dataset sources, annotation dimensions, and label definitions.

Module	Item	Description
Data sources	Schools	3 universities and 1 secondary school
Data sources	Courses	Comprehensive English, English reading, oral English, listening and speaking
Data sources	Teachers	12 English teachers
Data sources	Lessons	48 full lessons
Data sources	Duration	About 40 min per lesson; about 32 h in total
Data acquisition	Panorama video	Fixed rear-view classroom recording for whole-class interaction distribution
Data acquisition	Teacher close-up video	Front-facing teacher view for pose, gaze, and gesture observation
Data acquisition	Audio	Ambient microphone and lavalier microphone for classroom speech capture
Data acquisition	Text	ASR transcripts with manual correction
Segmentation	Unit	1 min per interaction segment
Segmentation	Total segments	About 1,920 segments
Annotation	Annotators	3 experts with backgrounds in English pedagogy or applied linguistics
Annotation	Scoring scale	1–5 points for each dimension, plus an overall score
Annotation	Label form	Continuous score and mapped 5-level category
Reliability	Agreement metrics	ICC and weighted kappa
Data split	Split strategy	Teacher-disjoint split
Data split	Train / Val / Test	32 lessons / 8 lessons / 8 lessons
Annotation dimension	Questioning quality	Degree of openness, progression, and cognitive prompting in teacher questioning
Annotation dimension	Feedback effectiveness	Extent to which teacher feedback clarifies, extends, scaffolds, or redirects learning
Annotation dimension	Participation breadth	Distribution of participation opportunities across students and seat areas
Annotation dimension	Emotional climate	Degree of encouragement, responsiveness, tension reduction, and interaction support
Annotation dimension	Target-language interaction density	Frequency, continuity, and functional load of English used in classroom interaction

As shown in Table 1, the dataset in this article not only contains multi-source classroom record information, but also refines the quality of English classroom interaction into five interpretable evaluation dimensions. Considering that the speech rate, classroom rhythm, and physical habits of the same teacher may be mistakenly identified as "high-quality interaction" cues by the model, this article adopts a data segmentation method based on teacher segmentation. Specifically, 32 lessons from 8 teachers were used for training, 8 lessons from 2 teachers were used for validation, and another 8 lessons from 2 teachers were used for testing. The three parts maintain a general balance in terms of academic stage, course type, and school source. This division makes the focus of model learning no longer on the teacher's personal style, but on a more transferable English classroom interaction mode.

2.2 Multimodal feature extraction and intelligent evaluation model

After completing classroom segment segmentation and five dimensional label annotation, this article further constructs a multimodal intelligent evaluation model for English classroom interaction quality. The input of the model is the visual segment, audio signal, and transcribed text of the i-th interaction segment, and the output is the segment level interaction quality score and classroom level comprehensive result. Unlike the approach of directly concatenating multi-source features, this article emphasizes the complementary relationship between the three modalities in classroom scenarios. Specifically, the visual branch is responsible for depicting action and participation distribution, the audio branch is responsible for reflecting rhythm and speech turns, and the text branch is responsible for modeling discourse information such as questioning, feedback, and target language use [16]. Firstly, independently encode the three modal inputs of the i-th classroom segment, as shown in formula (1).

$$h_i^{(m)} = E_m\left(x_i^{(m)}\right), \quad m \in \{\mathrm{v}, \mathrm{a}, \mathrm{t}\} \tag{1}$$

In formula (1), $x_i^{(m)}$ represents the original input of the i-th classroom segment on the m-th modality. $h_i^{(m)}$ represents the corresponding modal representation. $E_m(\cdot)$ represents the encoder for this modality. v, a and t correspond to three modalities: visual, audio, and text. The visual branch mainly extracts from classroom videos head pose, gaze direction, gesture frequency, teacher-student turn allocation And raising and frequency. The audio branch mainly extracts speech rate, pause ratio, pitch variation, and teacher student talk ratio. The text branch establishes representations around question type, feedback move, lexical diversity, L2 token ratio, and uptake pattern. Considering that the clues that best reflect the quality of interaction in different classroom segments are not consistent, this article further sets adaptive weights for the three modalities, as shown in formula (2).

$$\alpha_{\mathrm{i}}^{(\mathrm{m})} = \frac{\exp\left(w^{\mathrm{T}} h_i^{(m)}\right)}{\sum_{\mathrm{k}\in\{\mathrm{v,a,t}\}} \exp\left(w^{\mathrm{T}} h_i^{(k)}\right)} \tag{2}$$

In formula (2), $\alpha_i^{(m)}$ represents the attention weight of the m-th modality in the i-th segment. w represents a trainable weight vector. k represents the summation index, used to traverse three modalities. T represents transposition operation. This adaptive weight enables the model to automatically adjust the contribution of different modalities based on the information density of the current segment, rather than assuming in advance which is always more important: text, audio, or visual [17]. After obtaining the modal weights, the three modal representations are further fused into a unified vector, as shown in formula (3).

$$z_i = \sum_{m\in\{\mathrm{v,a,t}\}} \alpha_i^{(m)} h_i^{(m)} \tag{3}$$

In formula (3), z_i represents the cross modal fusion representation of the i-th classroom segment, which preserves key information at the three levels of behavior, voice and discourse at the same time, helping to reduce the deviation caused by the loss of a single mode or noise. In the segment level evaluation stage, this article uses a linear prediction head to output continuous quality scores $\hat{y}_i = W_s z_i + b_s$. Among them, $\hat{y}_i$ represents the predicted score of the i-th segment. W_s and b_s respectively represent the weight matrix and bias term of the prediction layer. In order to elevate the segment judgment to the level of the entire class, the

predicted results of all segments within the same classroom are averaged and summarized, as shown in formula (4).

$$Q_c = \frac{1}{N_c}\sum_{i=1}^{N_c} \hat{y}_i \tag{4}$$

In formula (4), Q_c represents the classroom level interaction quality score for lesson c. N_c represents the number of segments included in the classroom. By averaging and summarizing, the total score of the classroom is supported by the continuous interactive performance of the entire class, without relying on individual high or low activity segments, which is more in line with the overall requirements of classroom evaluation [18]. In the actual classroom evaluation process, both fine-grained scoring and grading are often required [19]. Therefore, in order to maintain the sensitivity of score prediction and the stability of level classification in the model, this paper adopts a joint objective function of regression and classification in parallel during the training stage, as shown in formula (5) [20].

$$\mathcal{L} = \lambda_1 \,\mathrm{MSE}(y_i, \hat{y}_i) + \lambda_2 \,\mathrm{CE}(c_i, \hat{c}_i) + \lambda_3 ||\Theta||_2^2 \tag{5}$$

In formula (5), $\mathcal{L}$ represents the total loss function, and y_i represents the true continuous score of the i-th segment. MSE stands for mean squared error term. CE(·) represents the cross entropy term. c_i represents the true level label. $\hat{c}_i$ represents the predicted level label. Θ represents all trainable parameters. λ_1, λ_2 and λ_3 respectively control the relative weights of the three loss components. From the perspective of module functionality, Visual branch is mainly responsible for posture, head orientation, frequency of raising hands, and seat area activity statistics. The Audio branch mainly captures rhythm changes, pause ratios, speaking epochs, and the proportion of teacher-student discourse. The Text branch focuses on the depth of questioning, types of feedback, and density of interaction in the target language. The three branches collaborate under a unified evaluation objective, and the final output of the model is an interactive quality judgment result that can be traced back to specific classroom evidence. In summary, the architecture diagram of the audio-visual text fusion model proposed by the research institute is shown in Figure 2.

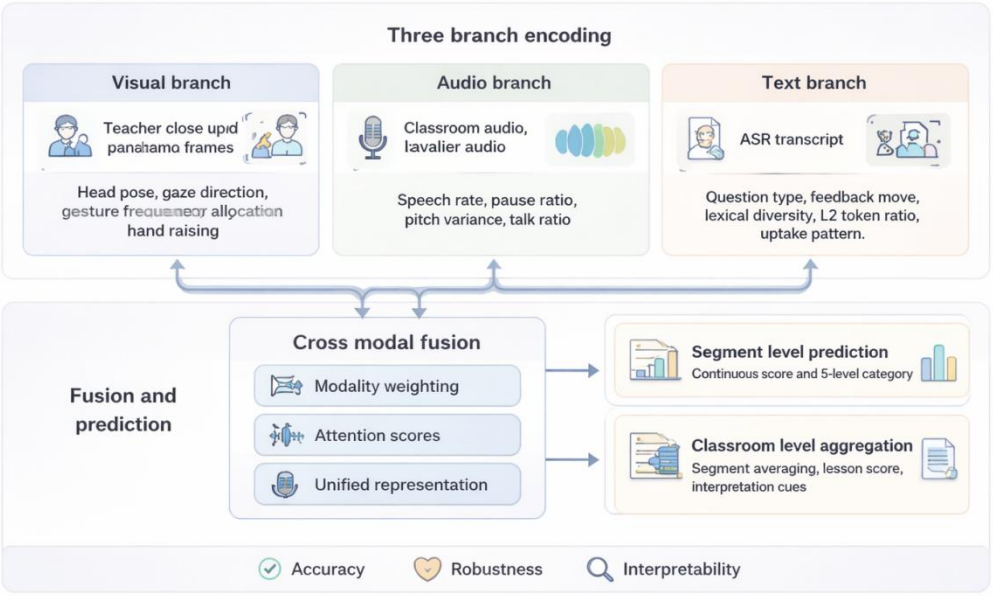


Figure 2: Architecture of the visual-audio-text fusion model.

As shown in Figure 2, the proposed audio-visual text fusion model consists of a visual branch, an audio branch, and a text branch. After encoding the three types of inputs separately, they enter the modal weight allocation module, which then fuses them into a unified representation and outputs fragment level interaction quality scores and level results. Finally, a comprehensive evaluation at the classroom level is obtained by aggregating fragments within the classroom.

2.3 Experimental settings and evaluation metrics

To test the practical effectiveness of the model in evaluating the quality of English classroom interaction, this article sets up experiments from three levels: comparative models, evaluation indicators, and robustness tests. All experiments were conducted based on the 1920 fragments mentioned above, with the training, validation, and testing sets divided according to the teacher dimension to avoid the same teacher appearing in different datasets at the same time, thereby reducing the model's dependence on individual teaching styles [21]. This classification method is consistent with the emphasis on real-world validation in recent years' research on intelligent classroom evaluation and multimodal learning analysis.

In terms of comparing model settings, this article did not stop at a single main model report, but instead constructed six sets of comparable baselines. The Text only model only uses classroom transcribed text and its derived discourse features to test the independent contributions of questioning, feedback, and target language usage information. The Audio only model only inputs speech rhythm, pause ratio, speaking duration, and teacher-student discourse ratio to observe the explanatory power of sound cues on interaction quality. The Visual only model only uses visual features such as head orientation, gaze distribution, gesture frequency, hand raising frequency, and seat area activity, which is consistent with the research path of teaching quality evaluation based on classroom video behavior recognition. Early fusion concatenates three types of modal features directly at the input and sends them to the prediction layer. Late fusion trains three unimodal branches separately, and then performs weighted ensemble on the output results. The proposed cross modal attention model is the main model in this article, which achieves cross modal fusion through adaptive modal weights to test the advantages of attention mechanisms in complex classroom segments. The existing research on English classroom interaction and classroom quality evaluation shows that classroom behavior, discourse and emotional cues often appear at the same time, and it is difficult to describe the interaction state stably only by a single mode.

The experimental training adopts a unified configuration: batch size is set to 16, maximum training rounds are 50, optimizer uses Adam, initial learning rate is set to 0.0001, and early stopping is triggered when there is no improvement in the validation set for 8 consecutive epochs. To control comparability between different models, all baselines and the main model use the same data partitioning, the same upper limit of training epochs, and the same evaluation process. This setting ensures both experimental reproducibility and allows performance differences to better reflect the model structure itself, rather than fluctuations in training conditions [22].

In terms of evaluation indicators, this article reports both classification indicators and regression indicators. Macro-F1 and Accuracy are used to evaluate the classification results of five level labels. Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are used to characterize the deviation between predicted scores and manually annotated scores [23]. Quadratic Weighted Kappa (QWK) is used to measure the degree of consistency between model output and expert ratings on an ordered scale. The Pearson Spearman correlation coefficient is used to examine the linear and rank correlations between classroom level comprehensive scores and manual classroom overall evaluations [24]. The reason for this

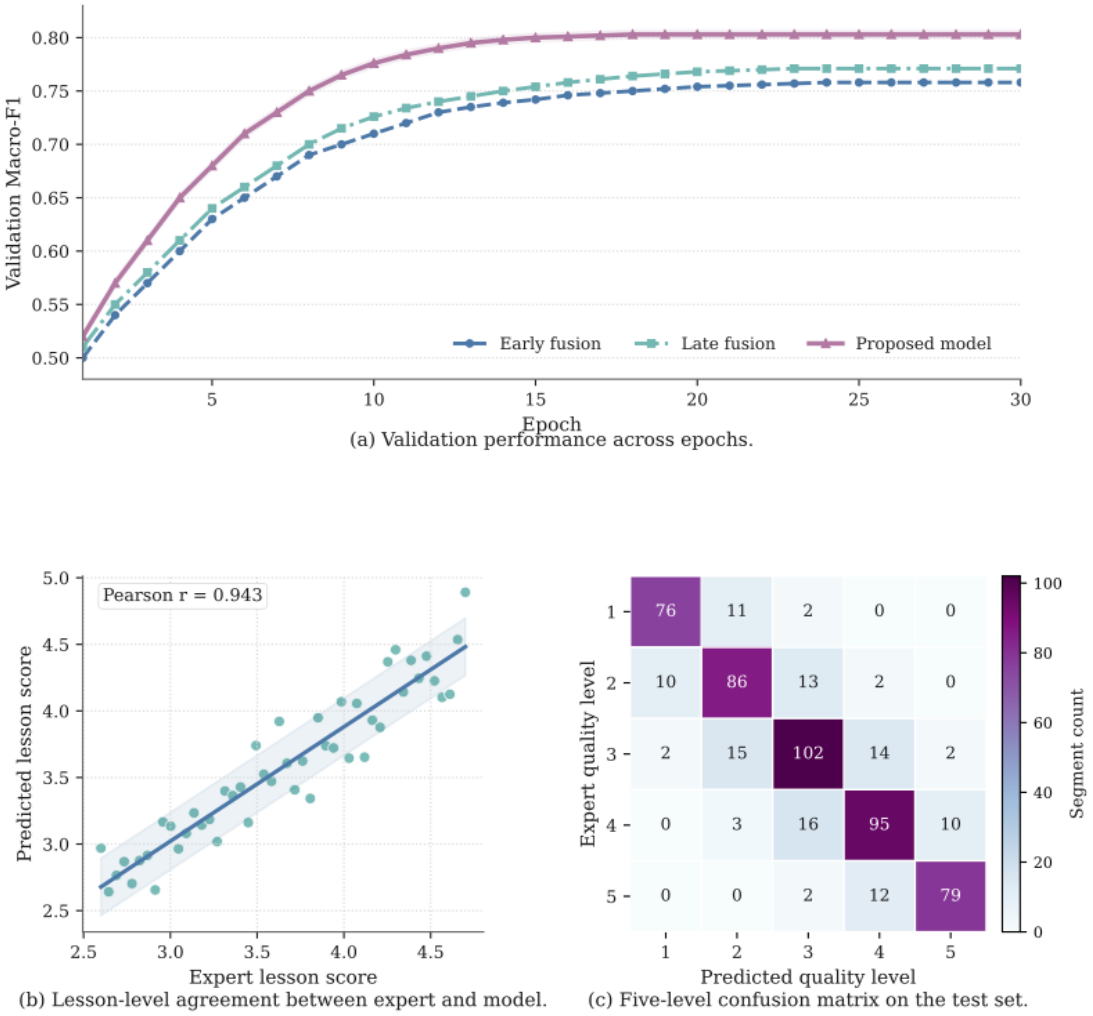
setting is that the quality of English classroom interaction has both a grading attribute and a continuous scoring attribute, and a single indicator is difficult to fully present the model performance. In recent years, multimodal teaching analysis research has increasingly emphasized that the results should not only be "correctly divided", but also consistent with real teaching judgments and have interpretability.

In the robustness testing section, this article conducts four additional types of tests. One is cross tutor validation, which focuses on examining the model's transferability on unseen teachers. The second is cross class type validation, which observes the adaptability of the model to changes in course types through cross testing between comprehensive English, reading classes, and speaking classes. The third is the missing modality test, which involves sequentially masking text, audio, or visual inputs during the testing phase to evaluate the extent of model degradation under modal loss conditions. The fourth is feature group ablation, which removes questioning features, feedback features, prosodic features, and participation distribution features to analyze the support strength of different feature clusters for the results. This type of setting corresponds to the focus on robustness, deployability, and teacher interpretable feedback in recent reviews of multimodal learning analysis and research on teacher decision support [25].

3 Results and Discussion

3.1 Overall performance and reliability analysis

To test the reliability of the five dimensional labeling system constructed in this article and the overall performance of the proposed model in the evaluation task of English classroom interaction quality, this study first reports the consistency of manual labeling, then compares the overall results of different models, and finally examines the consistency between classroom level prediction results and expert ratings. The convergence, classroom level fitting relationship, and five level label distribution of the model during the validation phase are shown in Figure 3.



(a) Validation performance across epochs.

(b) Lesson-level agreement between expert and model.

(c) Five-level confusion matrix on the test set.

Figure 3: Validation performance, lesson-level agreement, and five-level confusion patterns of the proposed model.

As shown in Figure 3 (a), the proposed model shows a faster increase in Macro-F1 during the validation phase and tends to stabilize around the 18th epoch, ultimately remaining around 0.803, higher than the early fusion of 0.758 and the late fusion of 0.771. This trend indicates that cross modal attention mechanisms can more effectively integrate visual, audio, and textual information, enabling the model to maintain a smoother convergence state in the later stages of training. The manual annotation results show that the within group correlation coefficient of the five dimensional total score reaches ICC = 0.86,questioning quality, feedback effectiveness, participation breadth, emotional climate The weighted Kappa values for target language interaction density are 0.74, 0.79, 0.76, 0.73, and 0.81, respectively. This indicates that the annotation system used in this article has high consistency and can provide stable supervision signals for subsequent model training. As shown in Figure 3 (b), there is a clear positive correlation between the total score of classroom level prediction and the total score of experts, with a Pearson correlation coefficient of 0.861. The scatter points are distributed along the fitted line, indicating that the model can not only identify interactive differences at the segment level, but also maintain good evaluation ranking at the whole class level. Figure 3 (c) further indicates that the errors of the five level labels are mainly concentrated between adjacent levels, especially between the second and third levels, and between the third and fourth levels, while there are fewer severe misjudgments spanning two

or more levels. This indicates that the model has been able to grasp the overall level of classroom interaction quality relatively stably, but there is still room for further optimization in identifying boundary segments. The overall differences in multiple evaluation indicators among different models are shown in Figure 4.

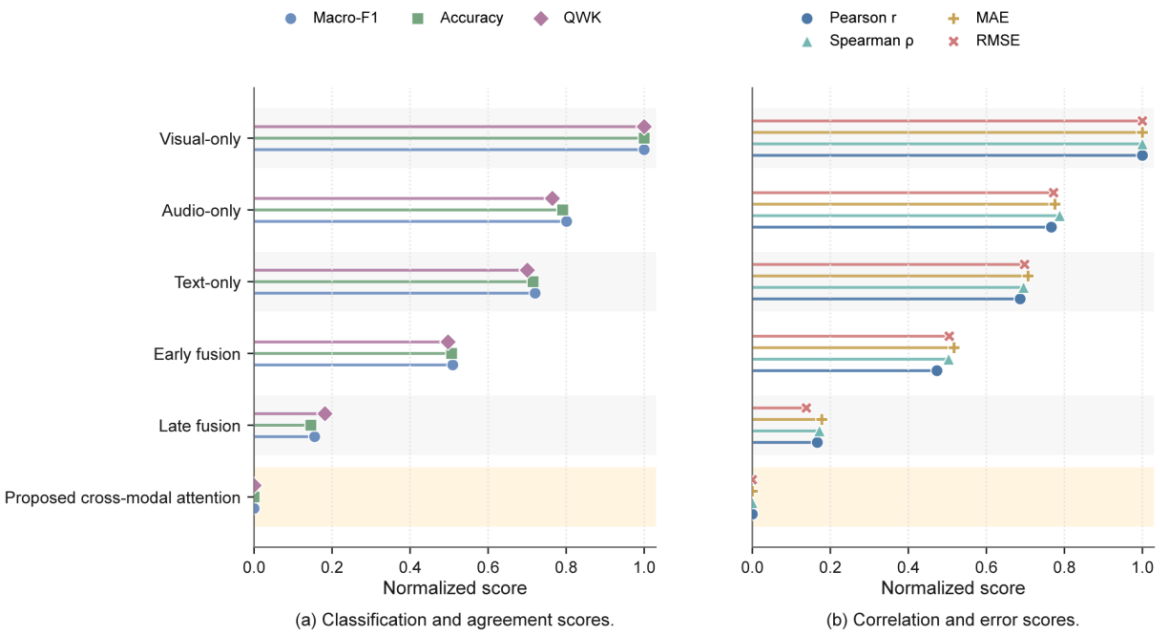


Figure 4: Multi-metric performance profiles of unimodal, conventional fusion, and proposed models on the test set.

As shown in Figure 4 (a), in terms of classification and level consistency related indicators, the proposed model is at the highest level in Macro-F1, Accuracy, and QWK, and its overall performance profile is significantly better than the three types of single modal models and two conventional fusion methods. This indicates that the proposed model has stronger advantages in five level label discrimination and ordered level consistency, and further validates the effectiveness of adaptive cross modal fusion in evaluating classroom interaction quality. At the same time, the text unimodal model performs the best among the three types of unimodal methods, indicating that questioning methods, feedback types, and target language usage information have high explanatory power for judging interaction quality. As shown in Figure 4 (b), the proposed model also maintains a leading position in classroom level correlation and error metrics. Its Pearson correlation coefficient is 0.861, Spearman correlation coefficient is 0.849, and it achieved the lowest values on the Mean Absolute Error and Root Mean Square Error indicators, which are 0.298 and 0.401, respectively. This result indicates that the proposed model can not only accurately distinguish the quality of classroom interaction at the level of grades, but also be closer to expert ratings in predicting continuous scores. Compared with early fusion and late fusion, the advantages of the proposed model are not limited to improving a single indicator, but are reflected in multiple aspects such as classification judgment, level consistency, classroom level relevance, and error control.

The results indicate that the five dimensional labeling system constructed in this paper has good manual consistency, and the proposed cross modal attention model exhibits strong stability and high evaluation consistency at both the fragment level and classroom level. Furthermore, there is a clear complementary relationship between visual, audio, and textual classroom evidence, and this complementary relationship can only be more fully utilized under adaptive weight allocation conditions. This discovery provides a more reliable basis for

subsequent ablation analysis and interpretability discussions.

3.2 Ablation study and cross-modal contribution analysis

To further illustrate the specific role of different modalities in evaluating the quality of English classroom interaction, this section analyzes from two aspects. Firstly, the degradation extent of the model performance after removing the single mode is investigated. Secondly, the dependence of different interaction dimensions on text, audio and visual evidence is compared. The verification process after modal removal and the changes in test set indicators are shown in Figure 5.

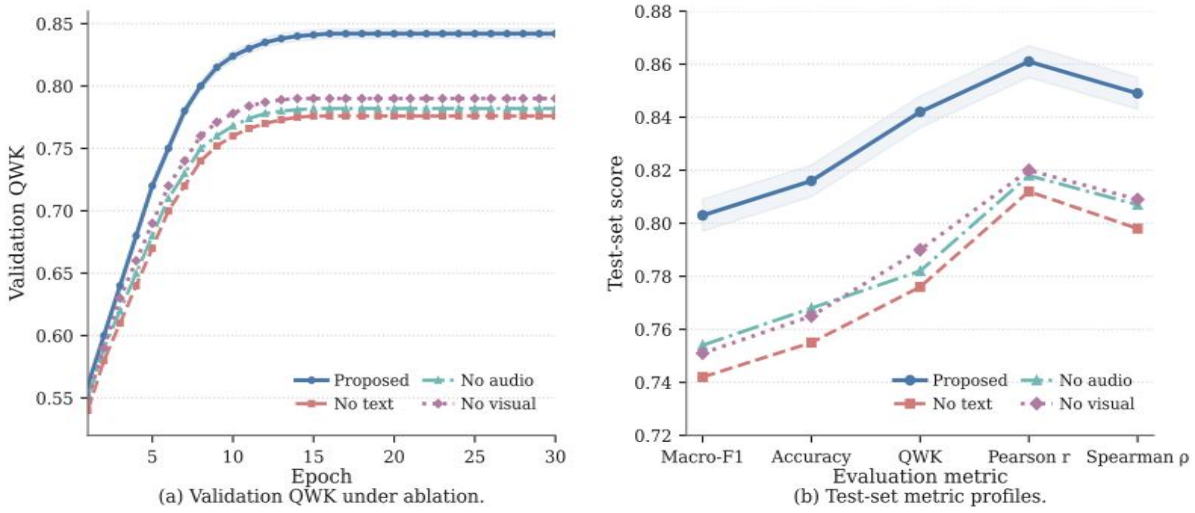
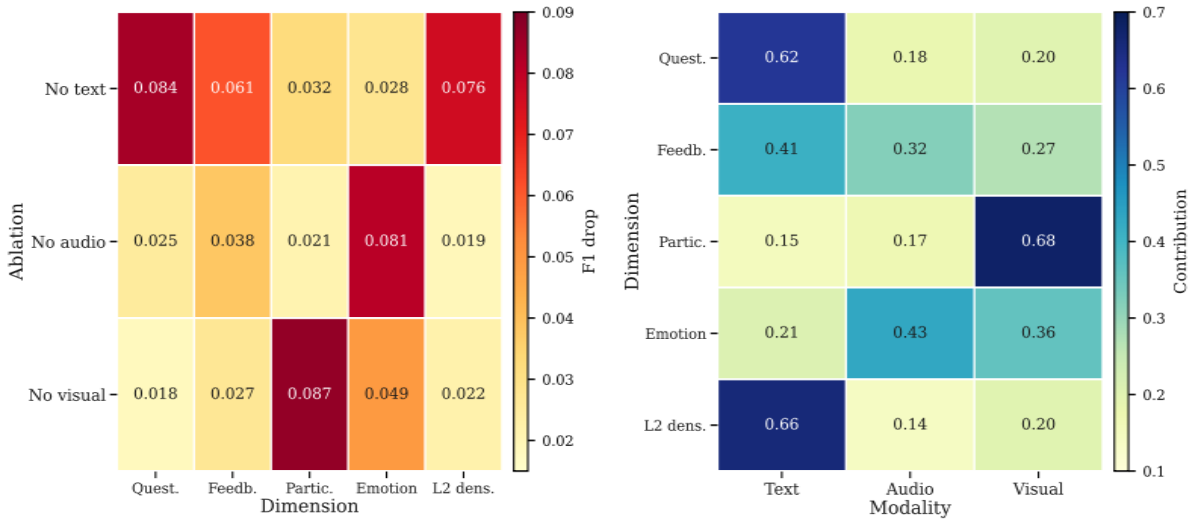


Figure 5: Validation degradation and test-set performance profiles under different modality ablation settings.

As shown in Figure 5 (a), the removal of any mode in the verification phase will lead to the decrease of the upper limit of convergence of QWK, of which the degradation after the removal of text is the most obvious. The validation set QWK of the complete model remained stable at 0.842, but after removing the text, it decreased to 0.776, a decrease of 0.066. After removing the audio, it decreased to 0.782, a decrease of 0.060. After removing vision, it decreased to 0.790, a decrease of 0.052. This result indicates that the text branch plays the strongest semantic support role in overall interaction quality judgment, while audio and visual information also provide irreplaceable supplementary evidence. As shown in Figure 5 (b), multiple indicators on the test set also exhibit the same trend. The Macro-F1, Accuracy, QWK, Pearson r, and Spearman correlation coefficients of the complete model are 0.803, 0.816, 0.842, 0.861, and 0.849, respectively. After removing the text, these five indicators decreased to 0.742, 0.755, 0.776, 0.812, and 0.798, respectively. After removing the audio, the values are 0.754, 0.768, 0.782, 0.818, and 0.807; After removing the visual, the values are 0.751, 0.765, 0.790, 0.820, and 0.809. From the overall trend, the loss caused by missing text is the greatest, followed by audio, and the impact after visual removal is slightly smaller, but still significantly lower than the complete model. This indicates that the benefits of cross modal fusion are not concentrated in a single indicator, but are reflected in the synchronous improvement of classification results, level consistency, and classroom level relevance. The performance degradation and modal contribution distribution in different dimensions are shown in Figure 6.



(a) F1 drops by ablation.

(b) Modality contributions.

Figure 6: Dimension-wise ablation effects and normalized modality contributions across interaction quality dimensions.

As shown in Figure 6 (a), the sensitivity of modal removal varies among different interaction dimensions. After removing the text, the F1 score of questioning quality decreased by 0.084, and the target language interaction density decreased by 0.076, indicating that these two dimensions mainly rely on textual evidence; After removing the audio, the F1 score of emotional climate decreased by 0.081, with the largest decrease, indicating that emotional atmosphere judgment relies more on speech intonation, pause organization, and speaking rhythm. After removing visual cues, the F1 score of participation breadth decreased by 0.087, indicating that participation coverage is mainly supported by behavioral cues such as raising hands, shifting gaze, and seat area activity. In contrast, feedback effectiveness decreased under all three types of modal conditions, with a decrease of 0.061 after removing text and 0.038 after removing audio, indicating that feedback effectiveness has a significant cross modal attribute. As shown in Figure 6 (b), further normalizing the modal contributions on different dimensions can provide a clearer view of the dominant sources of evidence for each dimension. The text contributions of questioning quality and target language interaction density reached 0.62 and 0.66, respectively, belonging to the obvious text dominated dimension. The visual contribution of participation breadth reaches 0.68, exhibiting prominent visual dominated features. The audio and visual contributions of emotional climate are 0.43 and 0.36, respectively, indicating that this dimension relies more on the joint action of audio visual cues. The contribution of feedback effectiveness in text, audio, and visual aspects is 0.41, 0.32, and 0.27, respectively, showing a relatively balanced cross modal distribution. From this, it can be seen that the quality of English classroom interaction is not determined by a single piece of evidence, but rather the result of the combined effects of language content, voice expression, and behavioral response.

The results indicate that the fundamental reason why cross modal fusion is superior to single modal models is that the evidence sources dependent on different interaction dimensions are not consistent. Text information is more suitable for depicting the depth of questioning and the use of target language, audio information is more suitable for reflecting emotional atmosphere and response rhythm, and visual information is more suitable for describing the scope of participation and behavioral distribution. When the model can dynamically call these complementary clues based on different segments, its performance in

overall evaluation will be more stable and closer to the composite structure of real classroom interaction.

3.3 Interpretable case analysis and pedagogical implications

To further illustrate the correspondence between model scoring and real classroom interaction evidence, this section selects a high-quality interaction segment and a low-quality interaction segment for case analysis, and combines the feature contribution results with additional teaching insights to verify which clues drive the model to give high scores, which clues drive the model to output lower scores, and whether these findings can be transformed into verifiable classroom improvement directions. The feature contribution results of high-quality and low-quality segments are shown in Figure 7.

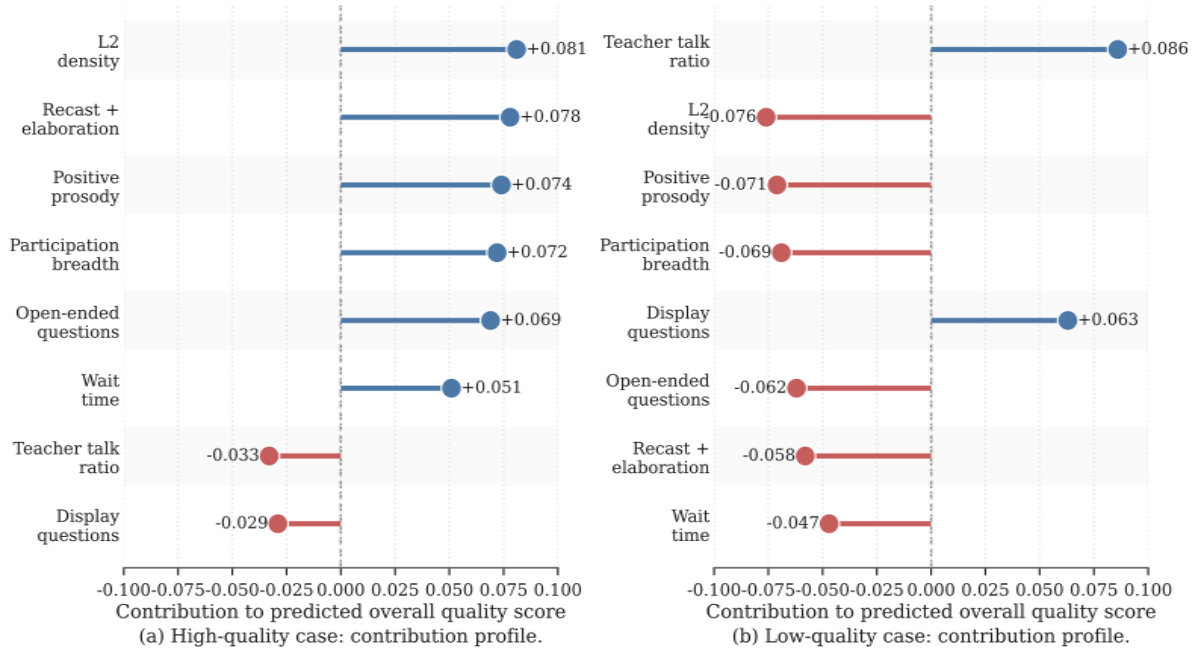


Figure 7: Feature contribution profiles in representative high- and low-quality classroom interaction segments.

As shown in Figure 7 (a), in high-quality interactive segments, the main basis for the model to give high score judgments is concentrated in open-ended questions, participation breadth, recast and elaboration, positive prosody And L2 density. Among them, the contribution of L2 density to overall prediction reaches+0.081, the contribution of replay and elaboration is+0.078, the contribution of positive procedure is+0.074, the contribution of participation breadth is+0.072, and the contribution of open-ended questions is+0.069. Relatively speaking, the contributions of teacher talk ratio and display questions in this segment are negative, with values of -0.033 and -0.029, respectively. This indicates that high-quality interactive segments receive high ratings mainly due to open questioning, broad participation coverage, feedback with extensibility, and more stable positive rhythms. As shown in Figure 7 (b), the contribution structure of low-quality interactive segments exhibits an opposite trend. In this segment, the teacher talk ratio and display questions become the main factors driving low-quality judgments, with their contribution values reaching+0.086 and+0.063, respectively. At the same time, open-ended questions, participation breadth, positive proof, and L2 density showed significant negative contributions to the prediction results, with values of -0.062, -0.069, -0.071, and -0.076, respectively. Based on classroom

records, it can be seen that the teacher's discourse in this segment is relatively high, with frequent use of demonstrative questions, short and evaluative feedback, and student responses concentrated on a few individuals. Therefore, the model accumulates more low-quality signals in the overall dimension. To verify whether these interpretable clues can further support teaching insights, the validation results of key variables such as open-ended questioning, waiting time, participation coverage, and student continuity in this study are shown in Figure 8.

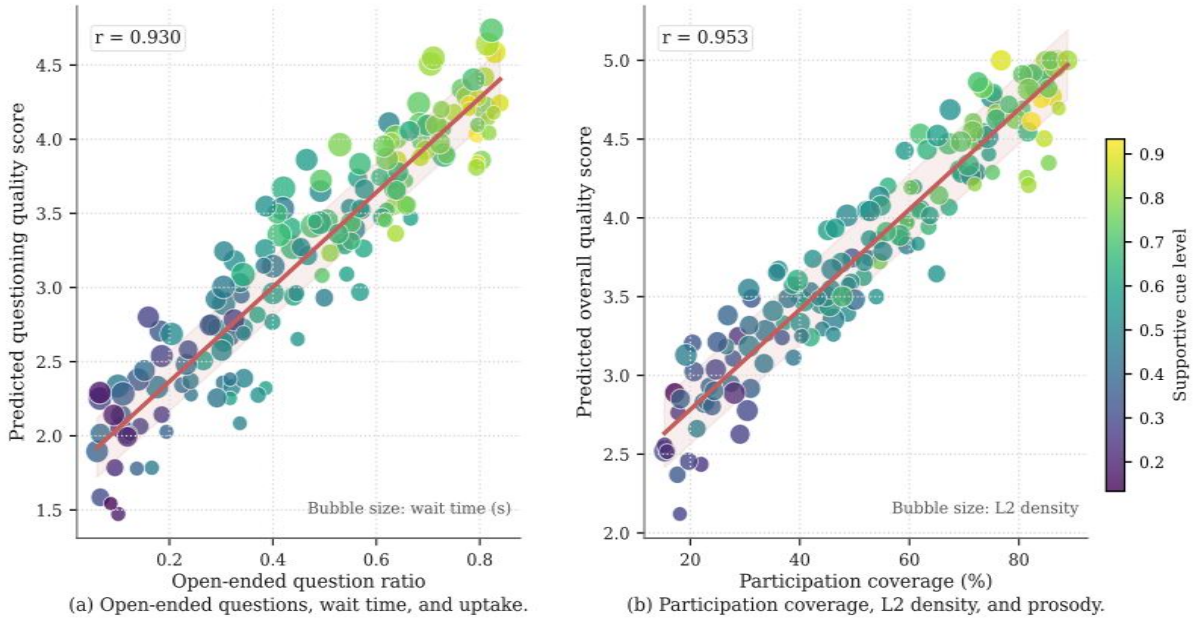


Figure 8: Data-supported validation of pedagogically actionable cues in classroom interaction quality prediction.

As shown in Figure 8 (a), the open-ended question ratio is significantly positively correlated with the predicted questioning quality score, with a correlation coefficient of r=0.835. In this relationship, bubble size represents wait time and color represents student uptake level. It can be seen that when the proportion of open-ended questions is higher, the waiting time is more sufficient, and students are able to continue speaking after feedback, the predicted score of questioning quality is significantly higher. This indicates that open-ended questioning is not an isolated variable, but rather a combination of conditions that are more conducive to high-quality interaction, including waiting time and student continuity. As shown in Figure 8 (b), there is also a stable positive correlation between participation coverage and the predicted overall quality score, with a correlation coefficient of r=0.860. In this figure, bubble size represents L2 density and color represents supportive proof level. With the increase of participation coverage, the overall classroom interaction quality score has significantly improved. Meanwhile, when the target language usage density is higher and the pronunciation and intonation are more supportive, the scatter is more concentrated in the upper right area. This indicates that the improvement of classroom interaction quality is often accompanied by broader student participation, more comprehensive use of target language, and more active voice expression.

This set of cases, together with the validation results, demonstrate that the model's high score judgment is mainly based on open-ended questioning, sufficient waiting time, wider student participation, feedback with scaffolding effect, and stable positive rhythm. Low scoring judgments are more likely to be triggered by a high proportion of teacher discourse,

frequent display questions, short feedback, and narrow participation scope. More importantly, the relationship analysis in Figure 8 indicates that these interpretable clues are not limited to empirical descriptions at the case level, but can be further validated in test data. From the perspective of teaching practice, these findings can be translated into clearer directions for improvement. Specifically, teachers can moderately increase referential/open questions, reserve more waiting time for students to respond, expand speaking coverage through follow-up, transfer, and roll call, and promote more obvious student uptake after feedback. For teacher training, such results can further implement the improvement of "interactive quality" at the level of observable and adjustable discourse behavior and classroom organization strategies.

4 Conclusion

This article focuses on the issues of strong subjectivity, insufficient process evidence, and limited feedback granularity in the evaluation of interactive quality in English classrooms. A multimodal intelligent evaluation framework that integrates video, audio, and classroom transcribed text has been constructed and validated from three levels: overall performance, modal contribution, and interpretable case studies. The main conclusions are as follows:

(1) The multimodal English classroom interaction quality evaluation framework constructed in this article can comprehensively characterize the core interactive features such as questioning, feedback, participation, emotional atmosphere, and target language use in the classroom, providing an operable technical path for joint evaluation at the segment level and classroom level.

(2) The experimental results show that the cross modal fusion model outperforms single modal methods in classification performance, level consistency, and classroom level relevance. There are significant differences in the dependence of text, audio, and visual evidence in different interaction dimensions, indicating that the quality of English classroom interaction itself has strong composite structural characteristics.

(3) The interpretability analysis further indicates that the model can not only achieve automatic scoring, but also identify key clues such as open questioning, waiting time, participation coverage, feedback extensibility, and target language usage intensity, providing more targeted references for teachers to improve classroom interaction.

However, the current sample size is still limited, and the coverage of classroom scenarios is not sufficient. The stability and adaptability of the model under real-time deployment conditions still need further testing.

Funding

This work was supported by the Jilin Vocational College of Industry and Technology Doctoral Research Initiation Fund Project (Project No. 25BS02SK) “Research on Curriculum Construction in Vocational College from the Perspective of Digitalized Education”

About the Author

Liu Yang, female, Lecturer, Ph.D., faculty member at Jilin Vocational College of Industry and Technology. The primary research areas include second language acquisition, second language motivation, vocabulary learning and teaching strategy research.

References

[1] Liu, Z., Sulaiman, T., & Che Nawi, N. R. (2025). Intelligent assessment of English teachers' classroom language interaction and emotional behaviour based on artificial intelligence. Scientific Reports, 15, 36010. DOI: 10.1038/s41598-025-20034-5.

[2] Wang, R., Chen, S., Tian, G., et al. (2024). Post-secondary classroom teaching quality evaluation using small object detection model. Scientific Reports, 14, 5816. DOI: 10.1038/s41598-024-56505-4.

[3] Zhou, Y., Zou, S., Liwang, M., Sun, Y., & Ni, W. (2025). A teaching quality evaluation framework for blended classroom modes with multi-domain heterogeneous data integration. Expert Systems with Applications, 289, 127884. DOI: 10.1016/j.eswa.2025.127884.

[4] Wang, S., & Dai, Y. (2025). Using lag sequential analysis to explore multimodal teacher-student interaction in EFL classrooms. Asian-Pacific Journal of Second and Foreign Language Education, 10, 37. DOI: 10.1186/s40862-025-00344-x.

[5] Querol-Julián, M. (2023). Multimodal interaction in English-medium instruction: How does a lecturer promote and enhance students' participation in a live online lecture? Journal of English for Academic Purposes, 61, 101207. DOI: 10.1016/j.jeap.2022.101207.

[6] Zhou, J., Li, C., & Cheng, Y. (2025). Transforming Pedagogical Practices and Teacher Identity Through Multimodal (Inter)action Analysis: A Case Study of Novice EFL Teachers in China. Behavioral Sciences, 15(8), 1050. DOI: 10.3390/bs15081050.

[7] Badem, F. (2025). Multimodal repair initiations in video-mediated EFL classroom interactions: Focus on screen-based and embodied actions. Learning, Culture and Social Interaction, 54, 100935. DOI: 10.1016/j.lcsi.2025.100935.

[8] Şimşek Tontuş, A., & Kuru Gönen, S. İ. (2025). Teachers' gestures in synchronous online language classrooms: embodied elicitation strategies for student participation. Social Semiotics, 1–23. DOI: 10.1080/10350330.2024.2448007.

[9] Mohammadi, M., Tajik, E., Martinez-Maldonado, R., Sadiq, S., Tomaszewski, W., & Khosravi, H. (2025). Artificial intelligence in multimodal learning analytics: A systematic literature review. Computers and Education: Artificial Intelligence, 8, 100426. DOI: 10.1016/j.caeai.2025.100426.

[10] Possaghi, I., Vesin, B., Zhang, F., et al. (2025). Integrating multi-modal learning analytics dashboard in K-12 education: insights for enhancing orchestration and teacher decision-making. Smart Learning Environments, 12, 53. DOI: 10.1186/s40561-025-00410-4.

[11] Sellberg, C., & Sharma, A. (2025). Toward multimodal learning analytics in simulation-based collaborative learning: A design ethnography of maritime training. International Journal of Computer-Supported Collaborative Learning, 20, 201–221. DOI: 10.1007/s11412-024-09435-2.

[12] Walkington, C., Nathan, M. J., Huang, W., Hunnicutt, J., & Washington, J. (2024). Multimodal analysis of interaction data from embodied education technologies. Educational Technology Research and Development, 72(5), 2565–2584. DOI: 10.1007/s11423-023-10254-9.

[13] Prinsloo, P., Slade, S., & Khalil, M. (2023). Multimodal learning analytics—In-between student privacy and encroachment: A systematic review. British Journal of Educational Technology, 54(6), 1566–1586. DOI: 10.1111/bjet.13373.

[14] Tavakoli, P. (2025). Assessment of second language fluency. Language Teaching, 58(3), 312–328. DOI: 10.1017/S0261444824000417.

[15] Tavakoli, P., Kendon, G., Muzhurnaya, S., & Ziomek, A. (2023). Assessing fluency in the Test of English for Educational Purposes. Language Testing, 40(3), 607–629. DOI: 10.1177/02655322231151384.

[16] Morrison, A., & Tavakoli, P. (2023). Task communicative function and oral fluency of L1 and L2 speakers. The Modern Language Journal, 107(4), 896–921. DOI: 10.1111/modl.12883.

[17] Yan, L., Wu, X., & Wang, Y. (2025). Student engagement assessment using multimodal deep learning. PLOS ONE, 20(6), e0325377. DOI: 10.1371/journal.pone.0325377.

[18] Das, R., & Dev, S. (2025). Optimizing student engagement detection using facial and behavioral features. Neural Computing and Applications, 37, 19063–19085. DOI: 10.1007/s00521-025-11317-z.

[19] Li, B., & Liu, P. (2024). Online Learning State Evaluation Method Based on Face Detection and Head Pose Estimation. Sensors, 24(5), 1365. DOI: 10.3390/s24051365.

[20] Chen, Y., Li, J., Liu, Y., & Jiang, F. (2025). A Novel Student Engagement Analysis of Real Classroom Teaching Using Unified Body Orientation Estimation. Sensors, 25(20), 6421. DOI: 10.3390/s25206421.

[21] Ding, X., & Mariano, V. Y. (2024). Learning Status Recognition Method Based on Facial Expressions in e-Learning. Journal of Advanced Computational Intelligence and Intelligent Informatics, 28(4), 793–804. DOI: 10.20965/jaciii.2024.p0793.

[22] Almulla, M. A. (2025). A multimodal emotion recognition system using deep convolution neural networks. Journal of Engineering Research, 13(2), 721–729. DOI: 10.1016/j.jer.2024.03.021.

[23] Sun, Z., Liu, H., Li, H., Li, Y., & Zhang, W. (2025). AVERFormer: End-to-end audio-visual emotion recognition transformer framework with balanced modal contributions. Digital Signal Processing, 161, 105081. DOI: 10.1016/j.dsp.2025. 105081.

[24] Ramaswamy, M. P. A., & Palaniswamy, S. (2024). Multimodal emotion recognition: A comprehensive review, trends, and challenges. WIREs Data Mining and Knowledge Discovery, 14(6), e1563. DOI: 10.1002/widm.1563.

[25] Guerrero-Sosa, J. D. T., Romero, F. P., Menéndez-Domínguez, V. H., Serrano-Guerrero, J., Montoro-Montarroso, A., & Olivas, J. A. (2025). A Comprehensive Review of Multimodal Analysis in Education. Applied Sciences, 15(11), 5896. DOI: 10.3390/app15115896.