scientific reports

Check for updates

OPEN

Intelligent assessment of English teachers' classroom language interaction and emotional behaviour based on artificial intelligence

Zhiming Liu[1,2], Tajularipin Sulaiman[2]✉ & Nur Raihan Che Nawi[2]

This study focuses on English teachers' classroom language expression, emotional changes, and teacher-student interaction behaviors, and proposes an intelligent evaluation model based on multimodal representation learning—Bimodal Modality Representation and Fusion Network (BMRN). The model integrates text, speech, and visual multimodal information, introduces a shared-private feature decoupling structure, and combines four types of losses (similarity, difference, cycle consistency, and reconstruction) to achieve cross-modal deep semantic fusion and accurate modeling. Experiments are conducted based on public datasets CMU- Multimodal Opinion-Level Sentiment Intensity (MOSI), Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI), and self-built teacher teaching data. Results show that BMRN outperforms existing advanced models in both classification and regression metrics. Specifically, on the CMU-MOSEI dataset, the mean absolute error is reduced to 0.547, and the Pearson correlation coefficient reaches 0.781; on the self-built dataset, the four-classification accuracy reaches 66.1%. Ablation experiments indicate that the text modality and cycle consistency loss contribute the most to model performance. Robustness tests verify the effectiveness of the cross-modal reconstruction mechanism in cases of modality missing. This model improves the ability of semantic understanding and emotion recognition for multimodal teaching behaviors, with strong generalization and practical application potential. This study provides strong technical support and data basis for teacher behavior evaluation and professional development in smart classrooms.

Keywords Artificial intelligence, Multimodal learning, Teacher evaluation, Binary modal representation learning, Cross-modal modeling

The deep integration of Artificial Intelligence (AI) into educational scenarios has driven innovations in research paradigms for classroom behavior. An English teacher's classroom is not a single process of knowledge transmission, but a multi-dimensional interactive system that integrates language communication, non-verbal interaction, and emotional perception[1–3]. Existing studies have shown that teachers' behavioral characteristics such as speech intonation and facial expressions not only affect students' emotional states but also significantly correlate with their learning motivation and cognitive load. For example, Paulmann and Weinstein (2025) examined the impact of improving primary school teacher trainees' voice awareness on their classroom communication through a voice training program. The results showed that after training, trainees spoke more slowly when conveying motivational intentions and reduced vocal effort, indicating that voice awareness training can significantly improve classroom communication and create a more supportive and autonomous learning environment for students[4]. Wang (2022) analyzed the impact of teachers' facial expressions on students' learning through three groups of video lectures (rich expressions, regular expressions, and audio-only). The results showed that teachers with rich expressions performed better in promoting students' social presence, arousal level, and long-term learning, and teachers' emotions could reduce students' cognitive load[5]. However, current classroom evaluations still rely heavily on methods such as manual observation scales and questionnaires. These

[1]Faculty of Foreign language Studies, Tianshui Normal University, Tianshui 741000, China. [2]Faculty of Educational Studies, Universiti Putra Malaysia, Serdang 43400, Malaysia. ✉email: tajulas@upm.edu.my

methods have limitations in assessing teachers' teaching language and emotional behaviors, including strong subjectivity, weak timeliness, and difficulty in dynamic tracking. They are particularly unable to capture the complex linkage characteristics of teachers' multimodal behaviors during teaching[6–8]. Therefore, there is an urgent need for an efficient, objective, and dynamically responsive multimodal evaluation mechanism to more accurately understand and provide feedback on teacher behaviors.

In recent years, the integrated analysis of modal data such as video, audio, and text has gradually become an important means for teacher teaching analysis, promoting the transformation of educational behavior research from static description to dynamic modeling[9,10]. Especially against the backdrop of the rapid development of multimodal perception technology, research on understanding teachers' behaviors by combining speech, facial expressions, and behavior sequences has continuously made progress. For example, Tanaka et al. (2024) evaluated the effectiveness of multimodal sensors (including accelerometers, gyroscopes, heart rate monitors, facial orientation trackers, and eye gaze trackers) in estimating students' concentration levels in online courses through experiments. They found that facial orientation and eye gaze had significant impacts in user-dependent and user-independent classifications, respectively, providing important insights for understanding and improving students' engagement in online learning environments[11]. However, current multimodal learning methods still face two key challenges. First, multimodal data exhibits strong heterogeneity, with differences in information distribution and semantics between different modalities[12–14]. Second, most existing methods focus on the concatenation or simple fusion of modal features, lacking effective modeling of shared information between modalities and modality-specific information, which easily introduces redundancy or blurs decision boundaries[15,16]. In actual teaching scenarios, English teachers' language intonation, dynamic expressions, and teaching content often evolve in a coordinated manner. How to extract stable and highly discriminative behavior representations from these multi-source information remains a technical difficulty that urgently needs to be overcome. Based on existing multimodal representation learning theories, this study proposes an innovative multimodal representation learning network that integrates a shared-private decoupling structure and a cross-modal interaction mechanism. It aims to address the complexity of teachers' classroom behaviors and the specific needs of educational scenarios. Different from traditional methods that only focus on modal fusion or simple feature sharing, this study emphasizes solving the ambiguity in teachers' behavior expression and redundancy in multi-source information in model design. For the first time, this study introduces the combination of Central Moment Discrepancy (CMD) and Hilbert-Schmidt Independence Criterion (HSIC) as structural regularizes, systematically optimizing the distribution consistency of the shared space and the independence of the private space, thereby enhancing the discriminative power and robustness of multimodal features. The cross-modal attention-based fusion unit designed in the model supports dynamic information transfer and strengthens in-depth semantic interaction between modalities, significantly improving the comprehensive understanding of teachers' classroom language, emotions, and interactive behaviors.

This model can not only be used to identify and evaluate English teachers' language styles, emotional states, and interaction patterns in the classroom, but also mine the deep structural characteristics behind teaching behaviors based on multimodal data. Thus, it can provide teachers with personalized suggestions for teaching improvement and analysis of ability development paths. Meanwhile, the evaluation results can assist schools or educational managers in optimizing teacher training content and enhancing classroom organization strategies. The behavioral representation information output by the model can also serve as an important reference for accurate student profiling, helping students obtain more adaptive teaching resources and situational support, and indirectly promoting the development of differentiation and precision in the teaching process. This study responds to the practical needs for objectivity, real-time performance, and in-depth understanding of classroom evaluation in intelligent education. It provides a basis for personalized teaching feedback, teaching style recognition, and teacher ability development evaluation, and offers references for the intellectualization of educational decision-making and classroom intervention strategies.

Literature review

With the continuous application of AI technology in educational scenarios, teaching behavior analysis is gradually shifting from qualitative observation to dynamic modeling based on multimodal data. Especially in the research on teachers' classroom behaviors, the integrated analysis of heterogeneous modal data such as speech, images, and text has become an important path to understand the teaching process and optimize teaching behaviors. Relevant studies on multimodal educational behavior modeling mainly focus on three directions: student behavior monitoring, teacher teaching feature perception, and multimodal fusion modeling methods.

In terms of student behavior monitoring, Watanabe et al. (2023) proposed using webcams for student engagement detection. They collected classroom behavior data through role-playing and achieved an 89.5% classification accuracy using a deep neural network model[17]. Subsequently, Watanabe et al. (2025) further used eye-tracking and a Long-Short-Term Memory (LSTM) model to predict students' self-reported comprehension levels, showing an F1 score of 0.886, which verified the effectiveness of temporal modeling[18]. Similarly, Trabelsi et al. (2023) developed a real-time visual monitoring system based on AI, which could identify students' emotions and attention states even when they were wearing masks, providing references for classroom regulation[19]. These studies indicate that multimodal perception technology can effectively capture changes in emotions and attention during learning behaviors, with good scalability.

In terms of teacher behavior perception, Ahuja et al. (2019) developed a real-time perception system that integrated visual and audio signals to support teachers' instructional feedback and professional growth[20]. Utami et al. (2019) systematically evaluated the applicability of facial expression recognition in teaching assessment from the perspective of image recognition, especially the model's robustness under the influence of lighting changes, occlusion, and head poses[21]. Prameela et al. (2024) designed an expression recognition model combining EfficientNet-B3 and a circle heuristic optimization algorithm, which achieved good recognition

results for students' cognitive emotions in low-resource environments[22]. These studies provide a technical basis for establishing teacher-cantered multimodal classroom behavior modeling. However, they generally focus on a single modality (such as vision) or specific tasks (such as expression recognition), making it difficult to cover the coupling characteristics between language, emotion, and interaction.

In terms of multimodal modeling methods, current mainstream approaches include modal concatenation, alignment, and collaborative modeling. Researchers have attempted to address the challenge of heterogeneity through modal alignment and shared-private space separation strategies. For example, Morita et al. (2025) combined Dual coding theory with generative AI to construct personalized reading materials, improving learning efficiency by 7.5%, which indirectly reflected the potential of semantic fusion between modalities[23]. In the field of educational management, Suryanarayana et al. (2024) emphasized that AI could be used to mine students' personalities and provide personalized teaching strategies[24]. Although these studies have promoted the development of educational intelligence, their modeling granularity is relatively coarse, lacking in-depth discussion on modal coupling mechanisms in teaching scenarios. There are still deficiencies in generalization ability and semantic consistency modeling in classroom multimodal interaction behaviors.

Despite progress in student monitoring and partial behavior recognition in current multimodal teaching behavior research, there remains a weakness in cross-modal joint modeling of teachers' classroom language, emotions, and interactive behaviors. In complex teaching contexts, different modalities often carry both shared and independent semantic information. However, existing studies generally overlook the structural decoupling of such shared and private features between modalities, leading to issues like information redundancy, feature interference, and semantic drift in models. This makes it difficult to achieve high-quality, interpretable behavior representation. Furthermore, traditional fusion methods mostly adopt static concatenation or shallow alignment strategies, failing to fully capture dynamic semantic interactions between multimodalities. Therefore, there is an urgent need to introduce more expressive mechanisms to construct deep semantic channels between modalities. To address this, this paper proposes a multimodal representation learning framework that combines shared-private structure decoupling and cross-modal attention mechanisms. Theoretically, this framework not only enables modeling of semantic consistency between modalities and independent modeling of modality privacy but also realizes dynamic transfer and in-depth interaction of multimodal semantics through attention mechanisms. This design effectively resolves the problem of redundant interference in multi-source heterogeneous information, enhances the discriminative power and robustness of behavioral features, provides a new structural paradigm in multimodal modeling methodology, and offers both theoretical and practical support for intelligent classroom evaluation and personalized teaching feedback.

Research model

Model overview and overall architecture

To achieve multimodal intelligent assessment of English teachers' language expression, emotional states, and interactive behaviors in the classroom, this study proposes an AI-based Bimodal Modality Representation and Fusion Network (BMRN). The input data of this model takes the sentence-level units of teachers' classroom speeches as the basic analysis granularity. For each sentence, the corresponding synchronous video clips, audio clips, and text content are aligned, unified modeled, and fused to capture complete and fine-grained emotional, semantic, and interactive features in each teaching language behavior. This approach not only ensures temporal consistency between multimodal inputs but also enhances the ability to accurately perceive teachers' classroom performance. Three types of modal information—text (Language, L), visual (Visual, V), and acoustic (Acoustic, A)—are extracted from the input data to respectively construct intra-modal temporal features. Then, semantic modeling is performed based on bimodal pairs to learn modal-shared and modal-private features, and deep fusion is achieved through an interaction mechanism, ultimately completing the intelligent assessment of teachers' emotional and behavioral states. Specifically, BMRN first encodes the three modalities in the input utterance into feature sequences $Z_l \in \mathbb{R}^{T_l \times d_l}$, $Z_v \in \mathbb{R}^{T_v \times d_v}$, and $Z_a \in \mathbb{R}^{T_a \times d_a}$. $T_{\{l,v,a\}}$ represents the number of time steps for each modality, and $d_{\{l,v,a\}}$ represents the modal feature dimension. Subsequently, the model sequentially performs shared-private feature modeling on the three bimodal combinations (L + V, L + A, V + A), integrates all semantic representations in the fusion stage, and outputs multimodal prediction results. The core design concept of BMRN lies in simultaneously modeling modal-shared features and modal-private features, thereby achieving a comprehensive construction of teachers' classroom behaviors across different perceptual channels. The overall structure of the model consists of three main components: a feature extraction module, a bimodal representation learning module, and a multimodal fusion module. The entire model structure is shown in Fig. 1.

In Fig. 1, the feature extraction module independently performs low-level feature modeling and temporal structure learning for each modality to capture intra-modal contextual associations. Subsequently, the bimodal representation learning module constructs modal-shared features and modal-private features for any pair of modalities. It introduces CMD loss and HSIC loss for optimization, respectively. Meanwhile, reconstruction loss and cycle consistency loss are added to enhance model robustness, and a cross-modal attention mechanism is introduced to model semantic interactions between shared and private features. In addition, a cross-modal attention mechanism is set within the module to model the interdependence between shared and private semantics, providing rich cross-modal semantic support for the final fusion stage. The multimodal fusion module aggregates shared and private features from different modality pairs and generates global representations through a structural fusion strategy, which are used for subsequent emotion recognition or behavior assessment tasks. This fusion not only covers the semantic complementary relationships between modality pairs but also retains modality-specific signal patterns and local feature expressions. Finally, the model outputs multi-dimensional evaluation vectors, supporting both classification and regression evaluation methods, to adapt to teacher classroom analysis tasks in various smart education scenarios.

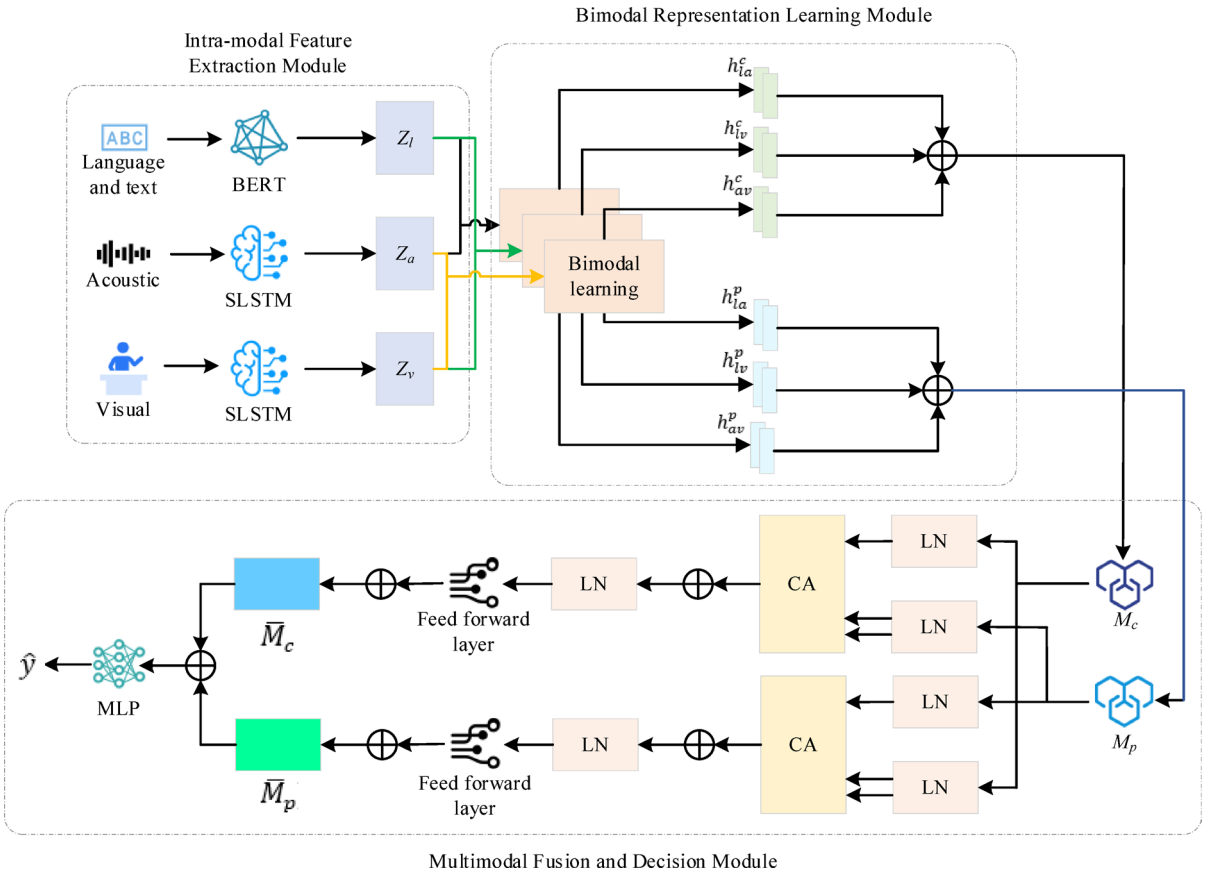

Fig. 1. BMRN model structure diagram.

Feature extraction module

In multimodal semantic assessment, the expressive power of intra-modal features directly affects the model's perception depth of teachers' classroom behaviors. Although text, audio, and visual modalities originate from the same classroom segment, they carry their own independent information structures and emotional cues[25,26]. To effectively extract high-dimensional semantic representations of these modalities, the study constructs a feature extraction module with a unified feature scale. It designs modality-specific pre-trained networks to ensure that the final features maintain consistency and complementarity in expression.

To ensure effective temporal alignment of text, audio, and visual modal data, unified time synchronization preprocessing is first performed on the multimodal raw data. Based on the video timestamps, text transcription content, audio signals, and video frame sequences are all segmented into corresponding time intervals, ensuring that the three-modal data within each time interval strictly correspond to the same classroom moment. In addition, to address the issue of different sampling frequencies across modalities, interpolation and resampling techniques are used to adjust the audio and visual modalities to the same time scale as the text time intervals.

For the text modality, the Bidirectional Encoder Representations from Transformers (BERT) pre-trained language model is adopted as the main feature extraction tool. BERT is trained on large-scale unsupervised text data and can capture rich linguistic semantics and emotional information[27]. Specifically, the first word vector from the last layer of the BERT model is used as the vector representation of the entire sentence, which contains comprehensive encoding of the whole sentence and reflects sentence-level semantic features. Let the input feature of the text modality be I_l, then the text representation vector Z_l can be expressed as:

$$Z_l = BERT(I_l; \gamma_l^{bert}) \in \mathbb{R}^{d_l} \tag{1}$$

γ_t^{bert} is the parameter of BERT and d_l is the text vector dimension.

For the speech and visual modalities, to comprehensively capture temporal dependencies and contextual information in sequences, a Stacked Bidirectional Long Short-Term Memory (SLSTM) is employed. The bidirectional structure allows the model to consider both past and future information when processing sequences, thereby generating more representative feature expressions[28]. The audio modality input I_a and visual modality input I_v are encoded separately, and the hidden state sequences are obtained using the stacked bidirectional LSTM. Finally, the terminal hidden state of the sequence is selected as the overall representation of the modality, denoted as Z_a and Z_v. Their mathematical expressions are:

$$Z_a = SLSTM(I_a; \gamma_a^{lstm}) \in \mathbb{R}^{d_a} \tag{2}$$

$$Z_v = SLSTM(I_v; \gamma_v^{lstm}) \in \mathbb{R}^{d_v} \tag{3}$$

γ_a^{lstm} and γ_v^{lstm} respectively represent the stacked bidirectional lstm parameters of speech and visual modes. d_a and d_v respectively represent the dimensions of the corresponding modal vectors.

To achieve unified processing of multimodal information, for each modality $m \in \{l, v, a\}$, its original representation vector $Z_m \in \mathbb{R}^{d_m}$ is mapped to a feature space with fixed dimensions, forming the final feature representation $Z_m \in \mathbb{R}^{d_k}$ for subsequent multimodal fusion and representation learning. This mapping process is implemented through a fully connected layer, ensuring effective alignment and interaction of all modalities within the same dimensional space.

Binary modal representation learning module

To address the inter-modal differences and information complementarity in multimodal data, a bimodal representation learning module is designed to specifically capture shared semantic features between two modalities and their respective unique private information. This module constructs shared and private subspaces in the feature space, and decouples and fuses bimodal features through a structured learning strategy. Figure 2 shows the overall framework of this module, reflecting the decomposition of modal features, their interaction, and the optimization mechanism with multiple constraints.

Modal feature representation

The unimodal features extracted by temporal models can capture long-term contextual information in sequences, but struggle to effectively address redundancy and differences between different modalities. To overcome this limitation, the module is designed with a parallel structure of a common encoder and private encoders to separately learn shared features between bimodalities and their respective private features. Specifically, input vectors Z_s and Z_t from two different modalities are mapped into two independent representation spaces. In this mapping process, the common encoder $E_c(\cdot\,; \gamma_c)$ is responsible for extracting shared features, aiming to eliminate distribution differences between modalities and achieve information fusion and sharing. Meanwhile, to preserve the unique semantic information of each modality, private encoders $E_p^s(\cdot\,; \gamma_p^s)$ and $E_p^t(\cdot\,; \gamma_p^t)$ are tasked with learning private features of modalities *s* and *t*, respectively. The formal expression of the shared feature representation is as follows:

$$h_{st}^c = E_c(Z_s; \gamma_c) \tag{4}$$

$$h_{ts}^c = E_c(Z_t; \gamma_c) \tag{5}$$

The formal expression of private feature representation is:

$$h_s^p = E_p^s(Z_s; \gamma_p^s) \tag{6}$$

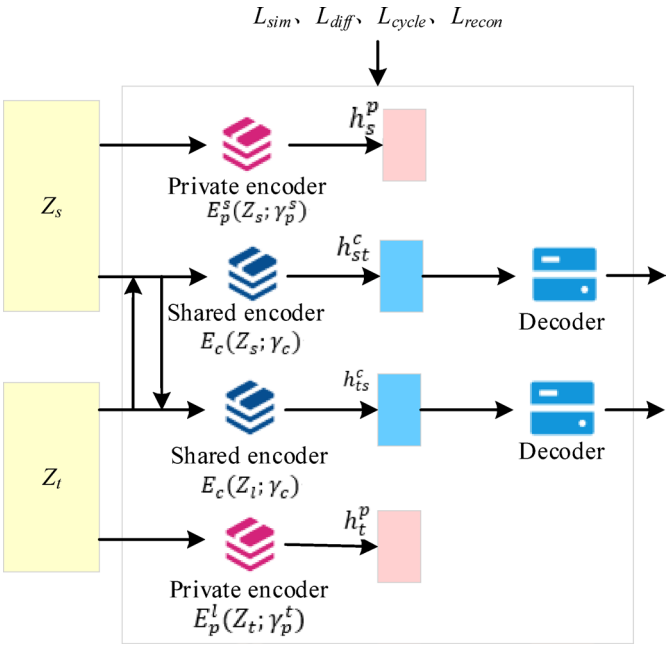

Fig. 2. Frame diagram of binary modal representation learning module.

$$h_t^p = E_p^l(Z_t; \gamma_p^t) \tag{7}$$

h_{st}^c and h_{ts}^c respectively represent the shared features of mode *s* and mode *t* extracted by a common encoder. h_s^p P and h_t^p represent the private features of the corresponding modes. All features are mapped to the same dimension space $\mathbb{R}^{d_k}$, which provides a unified representation basis for subsequent fusion and optimization. The common encoder parameter γ_c is shared between the two modes, while the private encoder parameters γ_p^s and γ_p^t are independently learned to ensure the effective distinction and expression of shared and private information.

Loss function

In the design of loss function, four structural constraints, such as similarity loss, difference loss, cyclic consistency loss and reconstruction loss, are introduced to capture the cooperative and independent characteristics of multi-modes.

Similarity loss In the process of multimodal fusion, effectively capturing shared semantics between modalities is crucial for enhancing representational capability. To this end, the model constructs a similarity loss term to guide the shared features of different modalities to converge in the latent space. This loss term measures the difference between the distribution of two modal representations from multiple dimensions, including mean and high-order central moments, based on CMD. CMD is highly sensitive to minor differences between distributions and can finely align representations of different modalities at the statistical structure level. For the shared representation vectors *X* and *Y* of two modalities, the formal definition of CMD is as follows:

$$CMD_k(X,Y) = \frac{1}{|b-a|}||\mathbb{E}[X] - \mathbb{E}[Y]||_2 + \sum_{k=2}^{\infty} \frac{1}{|b-a|^k}||c_k(X) - c_k(Y)||_2 \tag{8}$$

$\mathbb{E}[X]$ represents the expectation of vector *X*. $c_k(X) = \mathbb{E}\left[(x - \mathbb{E}(X))^k\right]$ represents the vector of central moment, and $[a,b]$ is the data interval. For any set of modes $s,t \in \{l,v,a\}$, the local CMD loss between vectors h_{st}^c and h_{ts}^c can be defined as:

$$L_{bisim} = CMD_k(h_{st}^c, h_{ts}^c) \tag{9}$$

Considering that there are three groups of multimodal combinations, namely L + V, L + A and V + A, this module models the losses of the three groups in a unified way, and obtains the similar losses among the overall shared representations:

$$L_{sim} = \frac{1}{3}\sum_{(s,t)\in\{(l,a),(l,v),(a,v)\}} L_{bisim} \tag{10}$$

In the optimization process, this loss term makes the shared codes of different modes close to the common semantics, and strengthens the information alignment ability across modes.

Difference loss Although shared features facilitate modal alignment, the unique information contained within each modality is equally critical. To encourage the model to fully preserve the exclusive semantics of each modality, a difference loss term is introduced to enhance the independence between shared and private representations. Specifically, the HSIC is adopted to measure the correlation between two sets of representations. This criterion constructs a covariance matrix using kernel methods and evaluates the statistical dependence between variables at the operator level. The definition of HSIC is as follows:

$$HSIC(X,Y) = \frac{1}{(n-1)^2} Tr(K_X H K_Y H) \tag{11}$$

K_X and K_Y are Gram matrices of samples in kernel space, with elements $K_{Xij} = k(X_i, X_j)$ and $K_{Yij} = k(Y_i, Y_j)$. H is a centralized matrix with:

$$H = I - \frac{1}{n}ee^T \tag{12}$$

I is the identity matrix, and *e* is a vector of ones with dimension *n*. In specific implementation, for each modality pair (s,t), the HSIC loss between the shared representation and its corresponding private representation is calculated respectively, and their average is taken as the local difference loss for the bimodal combination:

$$L_{bidiff} = \frac{1}{3}[HSIC(h_{st}^c, h_s^p) + HSIC(h_{ts}^c, h_t^p) + HSIC(h_s^p, h_t^p)] \tag{13}$$

The HSIC loss between private representations between each binary mode is:

$$L_{diff} = \frac{1}{3}\sum_{(s,t)\in\{(l,a),(l,v),(a,v)\}} L_{bidiff} \tag{14}$$

This design effectively avoids the information redundancy in shared and private space, provides a clearer feature decoupling mechanism for the model, and helps to improve the discrimination and generalization ability of the fusion representation.

Cycle consistency loss In multimodal representation learning, a single encoding-decoding process is often insufficient to ensure adequate information interaction between modalities. To enhance the stability and consistency of modal conversion, a cycle consistency mechanism is designed, enabling the model to achieve bidirectional information flow between modalities while maintaining consistent expression. Specifically, given the input vector Z_t of modality *t*, it is first mapped to the reconstructed vector Z'_s of the target modality *s* through the shared encoder E_c and cross-modal decoder D_c^s. Subsequently, this reconstructed result is used again as input, and restored to the initial modality through E_c and decoder D_c^t, generating Z''_t. The same process is applied to modality *s* to generate its cyclic output Z''_s. Throughout the process, the shared representation learned by the model must be able to complete the round-trip information mapping between modalities, ensuring the fidelity of semantic expression.

Based on the above process, the local cyclic consistency loss is measured by L1 norm to measure the deviation degree of samples before and after two reconstructions, and the form is as follows:

$$L_{bicycle} = \frac{1}{2}\left(||Z_t - Z''_t|| + ||Z_s - Z''_s||\right) \tag{15}$$

To ensure the equal weight of different modal combinations in the optimization process, the cyclic reconstruction errors among the three groups of modes are synthesized, and the global cyclic consistency loss is expressed as:

$$L_{cycle} = \frac{1}{3}\sum_{(s,t)\in\{(l,a),(l,v),(a,v)\}} L_{bicycle} \tag{16}$$

This loss term encourages the model to learn the common representation structure with good cross-modal recovery ability, which effectively improves the reversibility and consistency of the mapping between modes.

Reconstruction loss To enhance the expressive power of representations and suppress the generation of meaningless or oversimplified representations, a reconstruction loss is introduced to constrain the model's fidelity across various modalities. Within this framework, each modality pair (s,t) corresponds to a set of shared representations (h_{st}^c, h_{ts}^c) along with their respective private representations (h_s^p, h_t^p). These representations are transmitted to their corresponding cross-modal decoders D_c^s and D_c^t respectively, generating predicted vectors Z'_s and Z'_t for the original modal inputs. During the decoding process, the shared and private information act in conjunction, enabling the generated results to cover both the original modal details and cross-semantic features.

The local reconstruction error is measured by the Mean Squared Error (MSE), and the loss function is defined as follows:

$$L_{birecon} = \frac{1}{d_k}\left(\left|\left|Z_t - Z'_t\right|\right|_2^2 + ||Z_s - Z'_s||_2^2\right) \tag{17}$$

d_k is the dimension of the input vector, which is used to ensure the scale consistency of the error measure. Further integrate the reconstruction losses of all modal combinations and construct the overall reconstruction goal:

$$L_{recon} = \frac{1}{3}\sum_{(s,t)\in\{(l,a),(l,v),(a,v)\}} L_{birecon} \tag{18}$$

This loss design not only improves the restoration ability of the model to the input samples, but also helps to extract more robust and discriminating in-mode representation structures and enhance the generalization performance of the final fusion features.

Multimodal fusion module

In the complex teacher classroom scene, the information contained in language, vision and audio modes is not homogeneous, and the signals carried by them play an unequal role in the final evaluation task. Drawing on the "attention" mechanism in the Transformer structure, a Cross-modal Attention (CA) mechanism is introduced into the BMRN model to capture the dynamic dependencies between shared representations and private representations of different modalities. This enables selective alignment and transmission of information, as well as exploration of consistency and complementary features between modalities[29]. Set Query, Key and Value matrices in CA, which are defined as follows: given the input representation of mode α $X_\alpha \in \mathbb{R}^{T\times d_\alpha}$ and the input representation of mode β $X_\beta \in \mathbb{R}^{T\times d_\beta}$. Let the query matrix be $Q_\alpha = X_\alpha W_\alpha^Q$, and the key and value matrices be $K_\beta = X_\beta$ and $V_\beta = X_\beta W_\beta^V$. $W_\alpha^Q \in \mathbb{R}^{d_\alpha \times d_k}$ and $W_\beta^K, W_\beta^V \in \mathbb{R}^{d_\beta \times d_k}$ are learnable parameter matrices.CA output is:

$$Y_\alpha = CA_{\beta\to\alpha}(X_\alpha, X_\beta) = softmax\left(\frac{Q_\alpha K_\beta^T}{\sqrt{d_k}}\right)V_\beta \tag{19}$$

In practical implementation, to enhance the coverage capability of representations, the shared representations of six cross-modal combinations are stacked into a matrix $M_c = [h^c_{lv}, h^c_{vl}, h^c_{la}, h^c_{al}, h^c_{va}, h^c_{av}] \in \mathbb{R}^{6\times d_k}$, and the corresponding private representations are stacked into a matrix $M_p = [h^p_{lv}, h^p_{vl}, h^p_{la}, h^p_{al}, h^p_{va}, h^p_{av}] \in \mathbb{R}^{6\times d_k}$. M_c reflects the shared semantic information between modality pairs, while M_p captures the unique private information of each modality pair. This design ensures the orderly separation and expression of different modal features. After construction, information interaction is accomplished through the cross-modal attention mechanism, with M_c as queries and M_p as keys and values, to update the shared representations and promote adaptation to private features. The entire interaction process introduces residual connections based on the standard Layer Normalization (LN) operation, and connects to a feed-forward transformation layer for non-linear reconstruction. The finally improved representation is obtained as follows:

$$F_{c\to p} = softmax\left(\frac{Q_c K_p^T}{\sqrt{d_k}}\right) V_\beta + LN\left(M_c\right) \tag{20}$$

Further feed it into the feedforward network $f_\gamma(\cdot)$ to improve the representation ability and complete the cross-modal fusion:

$$\bar{M}_c = f_\gamma\left(LN\left(F_{c\to p}\right)\right) + LN\left(F_{c\to p}\right) \tag{21}$$

Finally, the enhanced shared representation $\bar{M}_c$ and the private representation $\bar{M}_p$ are spliced to form a joint representation vector:

$$h_{out} = [\bar{M}_c \oplus \bar{M}_p] \in \mathbb{R}^{12\times d_k} \tag{22}$$

The representation is mapped to a specific prediction task space through a Multilayer Perceptron (MLP):

$$\widehat{y} = MLP(h_{out}; \gamma_{out}) \tag{23}$$

This fusion method effectively enhances the model's ability to model cross-semantic information in multimodal inputs. It strengthens the dependency structure between representations. Thus, the model achieves stronger discriminative performance at the prediction level. By learning the dynamic relationships between different representations, the BMRN model ensures that the fusion results not only contain original modal information but also cover in-depth interactive semantics.

Model objective function and optimization strategy

To effectively evaluate the comprehensive performance of teachers' language use, interactive behaviors, and emotional expressions in English classrooms, the multimodal representation learning framework proposed in this study needs to optimize multiple objectives simultaneously during training. There are shared semantics and independent features among different modalities, and task requirements further demand the model to possess robustness and generalization ability. Therefore, the overall optimization strategy of the model revolves around an objective function composed of multiple loss functions, aiming to jointly model the semantic consistency, feature differences between different modalities, and the prediction accuracy of the final task. For classification scenarios such as teaching emotion category recognition, standard cross-entropy loss is adopted as the basic supervision signal:

$$L_{task} = -\frac{1}{n}\sum_{i=1}^{n} y_i \cdot \log\left(\widehat{y}_i\right) \tag{24}$$

y_i represents the true label of the *i*-th sample, and the prediction probability of the output of $\widehat{y}_i$ model. In regression tasks such as interactive intensity score prediction, MSE is used as the performance evaluation standard:

$$L_{task} = \frac{1}{n}\sum_{i=1}^{n} || y_i - \widehat{y}_i ||_2^2 \tag{25}$$

To prevent the model from focusing solely on target outputs, various structural auxiliary losses are introduced during training. Specifically, similar representations between modalities are guided to align via the similarity loss L_{sim}. Modality-specific representations are kept independent through the difference loss L_{diff}. The cycle consistency loss L_{cycle}) is used to enhance the ability of bidirectional semantic reconstruction between modalities. The reconstruction loss L_{recon} ensures that the semantic details of each modality can be effectively recovered, thereby improving the fidelity of the overall semantic representation. These multiple losses are weighted and combined into the final objective function:

$$L = L_{task} + \delta_1 L_{sim} + \delta_2 L_{diff} + \delta_3 L_{cycle} + \delta_4 L_{recon} \tag{26}$$

δ_1, δ_2, δ_3 and δ_4 are hyperparameters, which are used to adjust the weight balance of different loss items on the whole training process. Finally, the model carries out end-to-end training by minimizing the above objective function, and achieves a balance between multimodal feature learning and task prediction.

Experimental design and performance evaluation

Datasets collection

This study primarily adopts two public multimodal sentiment analysis datasets, CMU- Multimodal Opinion-Level Sentiment Intensity (MOSI) and CMU-Multimodal Opinion Sentiment and Emotion Intensity (MOSEI), as the main experimental basis to ensure the sufficiency of model training and the robustness of conclusions. CMU-MOSI contains 2,199 English short video samples, with each sample providing fine-grained scores for the intensity of opinion sentiment. CMU-MOSEI is one of the largest-scale datasets for multimodal sentiment and emotion recognition, covering over 23,500 samples. It includes six basic emotions and seven-level sentiment intensity scores, and is widely used in multimodal research tasks. Both datasets provide three types of modal information: video, audio, and text. They feature unified corpus structures and mature annotation standards, making them suitable as core data sources for model training and evaluation. Mean Absolute Error (MAE) and Pearson Correlation (Corr) are used to measure the fitting degree between model predictions and real labels. Acc2, Acc7, and F1 are adopted as sentiment classification metrics.

To further verify the model's adaptability in real teacher teaching scenarios, a small-scale sample set of English teachers' teaching videos is constructed for analyzing the model's generalization ability in educational fields. This dataset is sourced from the "Smart Primary and Secondary Schools" open platform, collecting classroom video clips of 12 English teachers. Through manual screening and editing, 400 teaching short videos (ranging from 2 to 10 s) with clear voices, stable images, and complete semantic expressions are retained, and language texts, audio features, and visual image frames are extracted synchronously. The data is annotated by three postgraduates with backgrounds in education and linguistics, all of whom received annotation training and had experience in foreign language education. Classroom assessment referred to three dimensions: clarity of language expression, naturalness of emotional regulation, and enthusiasm of teacher-student interaction, with four-level labels set as "Excellent", "Good", "Average", and "Need Improvement". Each video is scored independently by three annotators, and the final label is determined by majority voting. Among them, "Excellent" samples account for 24.5%, "Good" for 38.5%, "Average" for 26.0%, and "Need Improvement" for 11.0%. Cohen's Kappa is used to evaluate annotation consistency, with an average score of 0.74, indicating strong consistency in the annotations. These samples are divided into training set, test set, and validation set in an 8:1:1 ratio. All data collected in this study are teaching videos from public platforms, and all videos have undergone de-identification processing to ensure teachers' privacy and comply with relevant ethical norms and data usage requirements.

Experimental environment and parameters setting

The experimental environment and model parameters of this experiment are shown in Table 1.

To enhance the interpretability and classification accuracy of teaching quality analysis, the four levels of teachers' classroom performance, "Excellent", "Good", "Average", and "Need Improvement", are used for multi-classification modeling, defined as a 4-class task with Accuracy-4 and Macro-F1-4 as evaluation metrics. To adapt to the need for binary division of teacher evaluation in real educational scenarios, "Excellent" and "Good" are further classified into the high-performance category, while "Average" and "Needs Improvement" are classified into the low-performance category, forming a 2-class task with Accuracy-2 and Macro-F1-2 as metrics. This dual-task design aims to balance the continuity and hierarchical interpretability of teaching evaluation, enhancing the model's adaptability in various practical application scenarios.

In the experiment, a variety of classic and advanced multimodal fusion models are selected for comparison, covering both traditional machine learning methods and deep learning methods, to comprehensively evaluate the performance of the proposed model. The comparative models include: Early Fusion-LSTM (EF-LSTM),

Configuration/parameters	Detailed information
Operating system	Ubuntu 16.04.6 LTS
CPU	Intel Xeon E5-2620 12-core 24-thread 3.0 GHz
Graphics coprocessors	NVIDIA GeForce 2080Ti 1815 MHz 4352 stream processor
Programming language	Python 3.9
Deep learning framework	TensorFlow
Maximum length of text	256
Camcorder	16 kHz
Video frame rate	25fps
Number of multi-headed attention heads	8
δ_1, δ_2, δ_3. δ_4	0.5, 0.5, 0.3,0.2
Weight attenuation coefficient	0.01
Optimizer	AdamW
Dropout	0.3
Learning rate	0.0003
Epoch	20
Batch size	32
Activation function	ReLU

Table 1. Experimental environment and parameter settings.

which concatenates features from each modality and inputs them into an LSTM network for temporal modeling, representing the most basic early fusion approach. Tensor Fusion Network (TFN), which models full-order interactions between modalities through tensor outer products to capture complex fusion features. Low-rank Multimodal Fusion (LMF)[30], which achieves efficient multimodal feature fusion through low-rank tensor decomposition. Hybrid Contrastive Learning for Multimodal Data (HyCon)[31], which adopts a hybrid contrastive learning strategy to improve the consistency and discriminative power of inter-modal representations. Modality-Agnostic Gated BERT (MAG-BERT)[32], which uses a gating mechanism to flexibly integrate multimodal information, balancing modal robustness and dynamic adaptability of information fusion. These models are widely used and publicly validated in multimodal sentiment analysis and related tasks, with strong representativeness and reference value.

Performance evaluation

Model performance comparison

The performances of EF-LSTM, TFN, LMF, HyCon, MAG-BERT and the BMRN model proposed in this study on CMU-MOSI are shown in Fig. 3.

In Fig. 3, BMRN achieves the best performance in terms of classification accuracy Acc2 and F1 score, reaching 85.6% and 85.5% respectively, which are 0.47% higher than those of the current optimal HyCon model. In terms of regression metrics, BMRN has an MAE of 0.749, which is slightly higher than that of HyCon but still better than other baseline models such as MAG-BERT and LMF. In terms of correlation, BMRN reaches 0.78, which is basically the same as HyCon's 0.79 and higher than those of MAG-BERT and TFN. To further verify the significance of performance differences, 10-fold cross-validation and two-tailed paired t-test are used to evaluate the prediction results between each pair of models. The results show that BMRN has statistically significant differences compared with MAG-BERT and TFN in Acc2 and F1 metrics ($p < 0.01$). Although the difference with HyCon is small, BMRN maintains stability in multiple folds of training, showing good generalization ability. The performance of each model on CMU-MOSEI is shown in Fig. 4.

As Fig. 4 shows, BMRN achieves optimal or near-optimal performance in both classification and regression tasks. In terms of binary classification accuracy (Acc2), BMRN reaches 86.3%, which is 1.05% and 1.53% higher than that of HyCon and MAG-BERT, respectively, showing a significant improvement compared with the other models. In terms of F1 score, BMRN stands at 85.7%, slightly higher than that of HyCon and MAG-BERT, and performs more stably in balancing precision and recall. In the regression task, BMRN performs the best in MAE with a value of 0.547, significantly outperforming HyCon and MAG-BERT which also show good performance. Meanwhile, its Pearson correlation coefficient is 0.781, slightly lower than that of MAG-BERT but higher than that of HyCon and all early models. The results of the paired t-test show that BMRN has statistically significant differences compared with LMF, TFN, and EF-LSTM in terms of Acc2, F1, and MAE metrics ($p < 0.01$). Compared with MAG-BERT and HyCon, it also shows marginal significance in F1 and MAE ($p < 0.05$), indicating that the proposed model has better stability and discriminative ability in multiple dimensions. In summary, BMRN exhibits excellent multimodal fusion and representation capabilities on the CMU-MOSI and CMU-MOSEI datasets, enabling accurate and consistent judgments in large-scale real emotional and sentiment data.

The calculation costs of each model on CMU-MOSI and CMU-MOSEI datasets are compared, as shown in Table 2:

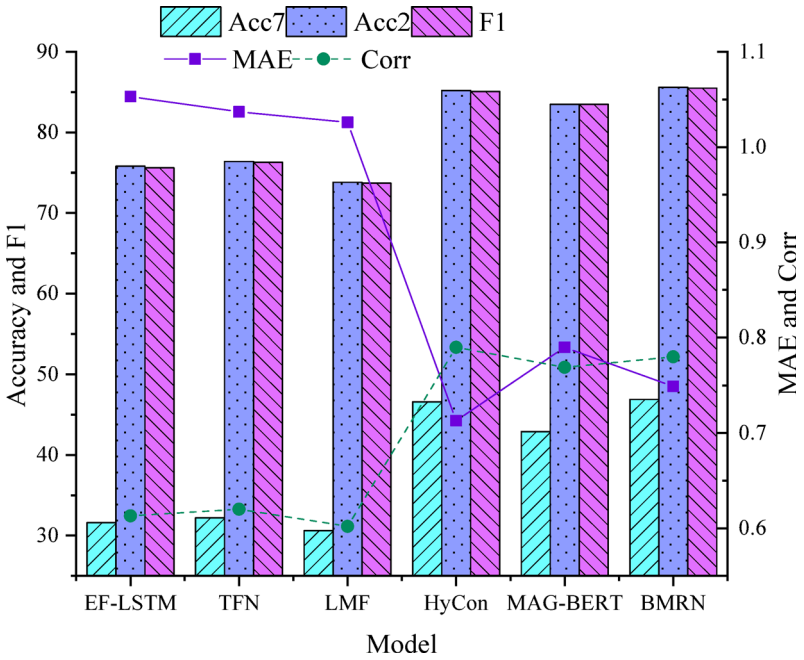

Fig. 3. Comparison of different models on CMU-MOSI.

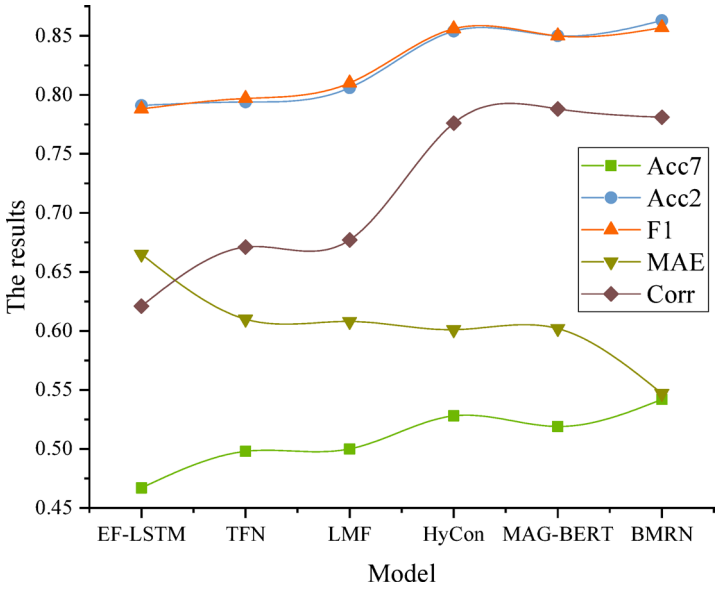

Fig. 4. Comparison of different models on CMU-MOSEI.

Model	CMU-MOSI training time (hours)	CMU-MOSI parameter scale (M)	CMU-MOSEI training time (hours)	CMU-MOSEI parameter scale (M)
EF-LSTM	2.1	3.5	4.5	3.5
TFN	3.8	5.2	7.9	5.2
LMF	3.2	4.8	7.1	4.8
HyCon	5.5	8.7	11.2	8.7
MAG-BERT	6.3	12.0	13.5	12.0
BMRN	5.8	11.5	12.7	11.5

Table 2. Comparison of calculation cost of each model.

From Table 2, the BMRN model has a slightly smaller parameter scale than MAG-BERT, but their training times are comparable, both being relatively time-consuming. Compared with traditional fusion models such as EF-LSTM and TFN, BMRN has higher computational costs, which is related to its adoption of a complex cross-modal attention mechanism and multi-loss collaborative optimization. Overall, while maintaining superior performance, BMRN's computational resource requirements are still within an acceptable range, and it has certain potential for practical application, especially in teaching platform deployment environments with strong computing power.

The performance of each model in the teacher teaching scenario is shown in Fig. 5:

In Fig. 5, the proposed BMRN model achieves the optimal performance across all four metrics. In the multi-classification task, BMRN attains an Accuracy-4 of 66.1% and a Macro-F1-4 of 65.6%, which are 4.59% and 4.63% higher than those of the sub-optimal model HyCon, respectively. In the binary classification task, BMRN leads the baseline models with an Accuracy-2 of 82.8% and a Macro-F1-2 of 82.3%, demonstrating stronger generalization and stability in distinguishing between high and low teaching performance. Overall, early-stage models generally achieve an accuracy of less than 60% in the four-class task, struggling to handle fine-grained teaching evaluation labels. Although MAG-BERT exhibits certain performance, it is limited by its reliance on semantic information in the teacher teaching dataset and is slightly inferior to BMRN in processing teacher-student interaction and emotional features. The results of the paired t-test show that BMRN has significant differences ($p < 0.01$) from TFN, EF-LSTM, and LMF in all four metrics. Compared with MAG-BERT and HyCon, BMRN has a marginally significant advantage ($p < 0.05$) in Macro-F1-4 and Macro-F1-2. In general, BMRN demonstrates better comprehensive performance in handling teaching quality evaluation tasks.

To further analyze the fine-grained classification performance of the model in the four-category task of teacher teaching evaluation, Table 3 shows the confusion matrix results of BMRN model on the self-built dataset test set, and reveals the main confusion relations between different categories.

In Table 3, the model exhibits a certain degree of confusion between "Excellent" and "Good" categories: it misclassifies "Excellent" as "Good" 9 times, accounting for nearly one-third of the total samples in the "Excellent" category. However, it demonstrates strong discriminative ability between "Need Improvement" and other categories, accurately identifying most "Needs Improvement" samples. Overall, the model is more prone to misclassification between adjacent levels, while showing robust recognition of extreme levels such as "Excellent" and "Need Improvement". The discriminative power between positive categories ("Excellent" and

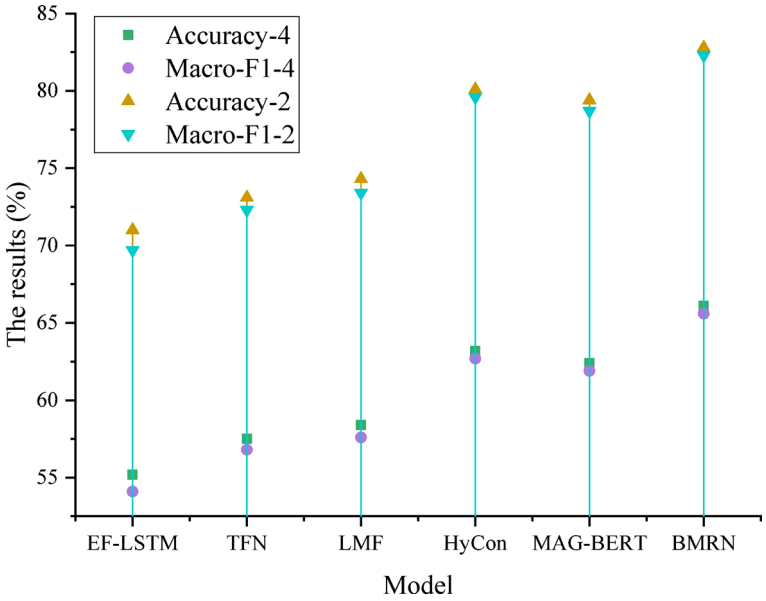

Fig. 5. The performance of each model on the self-built teacher teaching dataset.

Reality/prediction	Excellent	Good	Average	Need improvement
Excellent	19	9	2	0
Good	7	28	7	1
Average	2	5	17	6
Need improvement	0	1	3	8

Table 3. Four-category confusion matrix of BMRN model on teachers' teaching data test set.

"Good") is relatively limited, but the model can well distinguish the "Need Improvement" category. This provides a feasible basis for subsequently merging labels into three categories, such as "Excellent", "Average", and "Need Improvement".

Ablation experiment

In the ablation experiments, the text, audio, and visual modalities were removed separately to form models BMRN-l, BMRN-a, and BMRN-v, aiming to observe the impact of different modalities on performance. To quantitatively verify the importance of the four losses, one loss was removed each time by setting L_{sim}, L_{diff}, L_{cycle}, and L_{recon} to 0, resulting in models BMRN-sim, BMRN-diff, BMRN-cycle, and BMRN-recon. These models are used to analyze the effects on the expression of multimodal shared and private features, fusion performance, and final evaluation accuracy. The MAE and Corr results on the CMU-MOSI and CMU-MOSEI datasets are shown in Fig. 6.

In Fig. 6, in terms of modality ablation, the performance of BMRN-l (with the text modality removed) declines most significantly. Its MAE on the CMU-MOSI and CMU-MOSEI datasets rises to 0.827 and 0.827 respectively, while the Corr drops sharply to 0.237, indicating that text information remains the dominant modality in multimodal sentiment prediction tasks. In contrast, the performance degradation caused by removing the audio and visual modalities is relatively smaller. Their MAE values on CMU-MOSI are 0.568 and 0.57 respectively, and 0.568 and 0.57 on CMU-MOSEI. However, these values are still significantly lower than those of the complete model (0.749 and 0.547), which suggests that audio and visual modalities also play a non-negligible auxiliary role in enhancing the model's ability to perceive hierarchical and non-verbal emotional information.

In terms of loss term ablation, removing any loss function leads to performance degradation, verifying the key value of each loss in constructing the multimodal feature space. Specifically, removing the similarity loss L_{sim} and the difference loss L_{diff} causes the most significant damage to the model: on CMU-MOSI, the MAE rises to 0.83 and 0.831 respectively, while the Corr drops to 0.748 and 0.739 respectively. This indicates that the similarity loss plays a crucial role in constructing the shared feature space between different modalities, and the difference loss helps preserve the uniqueness of each modality to avoid feature redundancy. Removing the cycle consistency loss L_{cycle} and the reconstruction loss L_{recon} also causes performance degradation, though to a slightly lesser extent. BMRN-cycle achieves an MAE of 0.799 and a Corr of 0.758 on CMU-MOSI, suggesting that the cycle consistency mechanism helps maintain structural reversibility and semantic consistency of multimodal information in both private and shared spaces. BMRN-recon performs similarly, with slightly worse results than

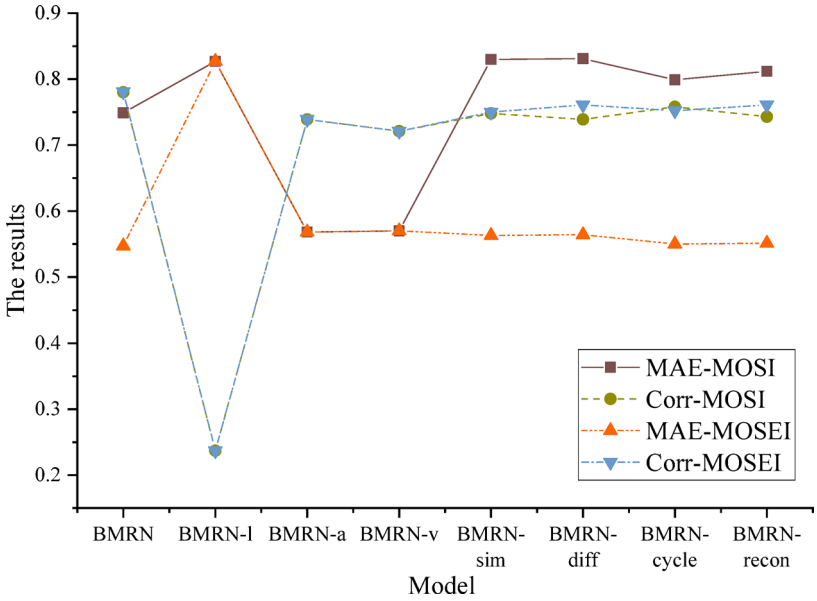

Fig. 6. Results of ablation experiment.

the complete model on both datasets, demonstrating that the reconstruction loss is significant for maintaining feature restoration and optimizing the quality of cross-modal alignment.

In summary, the results in Fig. 6 verify the dominant role of the text modality in sentiment prediction tasks, while emphasizing the indispensable collaborative role of audio and visual modalities in multimodal fusion. In addition, the four loss functions work synergistically, ensuring the effectiveness and stability of the BMRN model in multimodal fusion and representation learning from the dimensions of shared modeling, modal difference, cycle consistency, and semantic fidelity. The complete loss structure is the core support for achieving high model performance.

To further explore the model's robustness in the context of modality-missing scenarios, this experiment introduces a cross-modal completion mechanism. Specifically, for the missing modality input vector Z_t, the remaining modalities Z_s are retained. After unified encoding and cross-modal reconstruction, a substitute vector Z'_t is generated. Seven models are constructed: the complete input BMRN, and BMRN-t, BMRN-a, BMRN-v with text, audio, and visual modalities removed respectively, as well as BMRN-t-a, BMRN-t-v, and BMRN-v-a with two modalities missing. The experimental results show that the model's performance drops the most when text is missing, with the four-class accuracy rate falling to 49.63% and Macro-F1 slipping to 0.643. Among the two-modality-missing models, BMRN-t-a performs the weakest, with a four-class accuracy rate and Macro-F1 of only 46.27% and 0.612 respectively, indicating that the absence of linguistic cues has a significant impact on emotional judgment. In contrast, the model still maintains relative stability when audio and visual modalities are missing, verifying the effective support of the cross-modal reconstruction mechanism in enhancing model adaptability. The specific results are shown in Fig. 7.

Discussion

In this study, BMRN model shows excellent multi-modal integration ability, especially in teacher's emotion recognition and classroom behavior evaluation, which significantly surpasses the traditional model. This finding is consistent with previous research. Wang et al. (2024) showed that multimodal fusion models can effectively enhance the accuracy of sentiment analysis and behavior prediction[33]. Compared with single-modality research results, the proposed multimodal fusion strategy is evidently more adaptable to the dynamic classroom environment. The model's superior performance in multiple regression metrics such as MSE, MAE, and correlation reflects the in-depth mining and expressive power of multimodal information in teaching assessment. In classification metrics, the MSA-Class model also outperforms comparison methods. The significant increase in accuracy indicates that the model has unique advantages in fine-grained teaching behavior recognition and emotion state classification. This is in line with the multimodal analysis method proposed by Ezzameli and Mahersia (2023). They also found that multimodal models could better capture the subtle connections between emotions and behaviors, especially in scenarios with significant emotional fluctuations, where the model's sensitivity and accuracy are enhanced[34]. The results of the ablation experiment also provided in-depth insights into the model. When the text modality is removed, the model's performance drops significantly, verifying the core role of linguistic information in classroom assessment. This result is consistent with Zhao et al. (2023)'s research on emotions and language expression, who pointed out that the language modality was crucial for emotion judgment[35]. The absence of visual and audio modalities also leads to varying degrees of performance degradation, further emphasizing the synergistic effect of multimodal information. These findings not only validate the importance of multimodal fusion in teacher emotion and behavior assessment but also provide new ideas and methods for future educational data analysis.

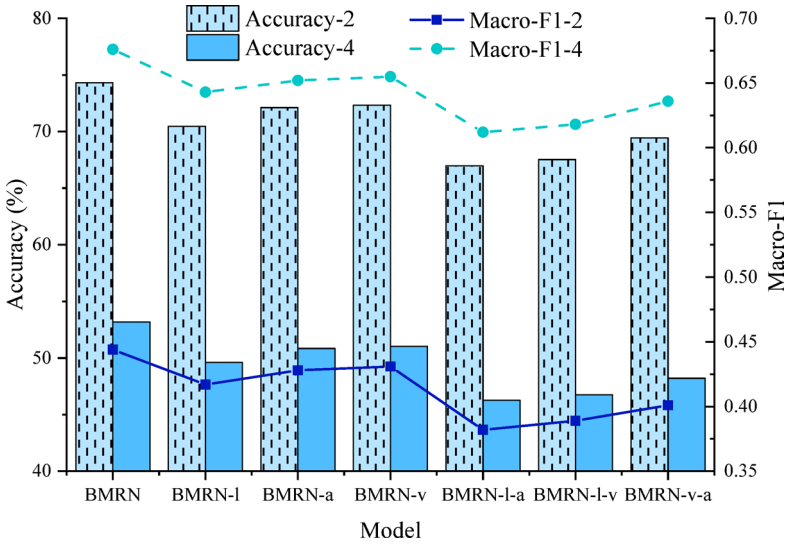

Fig. 7. Comparison of results when different modes are missing.

Conclusion

Research contribution

This study addresses the complex task of multimodal teaching behavior understanding and proposes the BMRN model, which centres on shared-private structure decoupling and cross-modal attention mechanisms. It achieves deep alignment between modalities and semantically consistent representation through collaborative optimization of multiple losses. The main contributions are as follows:

(1) Aiming at the coupling relationship between heterogeneous behavioral signals such as teachers' language, emotions, and interactions in teaching scenarios, a shared-private feature decoupling structure is introduced. This structure effectively separates shared semantics and unique features between modalities, improving the accuracy and expressive ability of cross-modal information modeling. Meanwhile, by introducing four loss functions, similarity loss, difference loss, cycle consistency loss, and reconstruction loss, structural breakthroughs are achieved in terms of semantic consistency, feature fidelity, and redundancy control.
(2) BMRN outperforms existing optimal models such as HyCon and MAG-BERT on key metrics including classification accuracy, F1 score, and MAE on the two standard datasets CMU-MOSI and CMU-MOSEI, demonstrating strong discriminative ability and semantic understanding ability. Especially on the self-built dataset for teacher teaching, BMRN achieves 66.1% in Accuracy-4 and 65.6% in Macro-F1-4 in the four-classification task, significantly surpassing other baseline methods, reflecting its strong generalization in complex educational scenarios.
(3) Ablation experiments show that the text modality is the dominant information source for emotion recognition, while audio and visual modalities provide important auxiliary signals, which indicates that the model reasonably models the weights of different modalities. In addition, the four loss functions play a synergistic role in feature decomposition and semantic preservation, which is a key guarantee for the performance improvement of BMRN. Further, in the modality missing experiment, the cross-modal reconstruction mechanism significantly alleviates the performance degradation caused by information loss, showing the model's robustness and adaptability in real teaching environments.
(4) The BMRN model not only verifies its strong performance on multiple general sentiment analysis datasets but also shows good practical application prospects in the self-built teaching behavior evaluation task. Its structural design is adapted to the multimodal data input of intelligent classroom platforms, and can effectively support functions such as teacher behavior evaluation, classroom feedback, and teaching optimization. It provides a feasible algorithm basis and methodological support for building an intelligent and interpretable teacher evaluation system.

In conclusion, through model structure innovation and system performance verification, BMRN provides a new research paradigm for multimodal teaching behavior modeling, and has important theoretical and practical significance in promoting the in-depth application of educational AI technology in actual classroom scenarios.

Future works and research limitations

This study still has the problems of limited data scale and limited situational adaptability, and the generalization ability of the model to cross-cultural classrooms and different teaching styles needs to be verified. In the future, multi-source heterogeneous data coverage will be expanded, and fine-grained modes such as eye movements and physiological signals will be integrated to enhance the richness and timeliness of evaluation dimensions. At the same time, causal reasoning and self-monitoring mechanism are introduced to explore a more explanatory

and transferable intelligent evaluation framework to serve diversified teaching scenarios and personalized teacher development needs.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author Tajularipin Sulaiman on reasonable request via e-mail Lzmai@163.com.

Received: 17 May 2025; Accepted: 11 September 2025
Published online: 15 October 2025

References

1. Daumiller, M. et al. Teaching quality in higher education: agreement between teacher self-reports and student evaluations. *Eur. J. Psychol. Assess.* **39** (3), 176 (2023).
2. Shabani Varaki, B. & Hosseingholizadeh, R. Evaluation of college teaching qualities. *Q. J. Res. Plann. High. Educ.* **12** (1), 1–21 (2023).
3. Al Maktoum, S. B. & Al Kaabi, A. M. Exploring teachers' experiences within the teacher evaluation process: A qualitative multi-case study. *Cogent Educ.* **11** (1), 2287931 (2024).
4. Paulmann, S. & Weinstein, N. Motivating tones to enhance education: the effects of vocal awareness on teachers' voices. *Br. J. Educ. Psychol.* **95** (2), 551–564 (2025).
5. Wang, Y. To be expressive or not: the role of teachers' emotions in students' learning. *Front. Psychol.* **12**, 737310 (2022).
6. Chen, X. & Chen, R. Examining the influence of the performance of mobile devices on teaching quality based on the technological acceptance mode. *Libr. Hi Tech.* **42** (2), 670–695 (2024).
7. Wang, H. Teaching quality monitoring and evaluation using 6G internet of things communication and data mining. *Int. J. Syst. Assur. Eng. Manage.* **14** (1), 120–127 (2023).
8. Almufarreh, A., Noaman, K. M. & Saeed, M. N. Academic teaching quality framework and performance evaluation using machine learning. *Appl. Sci.* **13** (5), 3121 (2023).
9. Wu, Y., Daoudi, M. & Amad, A. Transformer-based self-supervised multimodal representation learning for wearable emotion recognition. *IEEE Trans. Affect. Comput.* **15** (1), 157–172 (2023).
10. Manzoor, M. A. et al. Multimodality representation learning: A survey on evolution, pretraining and its applications. *ACM Trans. Multimedia Comput. Commun. Appl.* **20** (3), 1–34 (2023).
11. Tanaka, N. et al. Concentration estimation in online video lecture using multimodal sensors. Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing. 71–75. (2024).
12. Jiao, T. et al. a comprehensive survey on deep learning multi-modal fusion: methods, technologies and applications. Computers, Materials & Continua. **80**(1), (2024).
13. Shou, Y. et al. A comprehensive survey on multi-modal conversational emotion recognition with deep learning. arXiv preprint arXiv:2312.05735. (2023).
14. Du, B. et al. Heterogeneous structural responses recovery based on multi-modal deep learning. *Struct. Health Monit.* **22** (2), 799–813 (2023).
15. Wang, Q. et al. BiMSA: multimodal sentiment analysis based on BiGRU and bidirectional interactive Attention. *Eur. J. Artif. Intell.* **38** (3), 296–308 (2025).
16. Liu, Y. et al. Part-whole relational fusion towards multi-modal scene understanding. *Int. J. Comput. Vision*. **133** (7), 4483–4503 (2025).
17. Watanabe, K. et al. Engauge: engagement gauge of meeting participants estimated by facial expression and deep neural network. *IEEE Access.* **11**, 52886–52898 (2023).
18. Watanabe, K. et al. EyeUnderstand: dashboard for gaze and deep-learning driven comprehension estimation in online lectures. *IEEE Access.* **13**, 102220–202233 (2025).
19. Trabelsi, Z. et al. Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition. *Big Data Cogn. Comput.* **7**(1), 48 (2023).
20. Ahuja, K. et al. EduSense: Practical classroom sensing at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies.* **3**(3): 1–26. (2019).
21. Utami, P., Hartanto, R. & Soesanti, I. A study on facial expression recognition in assessing teaching skills: datasets and methods. *Procedia Comput. Sci.* **161**, 544–552 (2019).
22. Prameela, N., Das, M. S. & Borreo, R. D. Multi-head network based students behaviour prediction with feedback generation for enhancing classroom engagement and teaching effectiveness. *Int. J. Inform. Technol. Comput. Sci. (IJITCS)* **16**(5), 81–100 (2024).
23. Morita, R. et al. GenAIReading: augmenting human cognition with interactive digital textbooks using large Language models and image generation models. arXiv preprint arXiv:2503.07463. (2025).
24. Suryanarayana, K. S. et al. Artificial intelligence enhanced digital learning for the sustainability of education management system. *J. High. Technol. Manage. Res.* **35**(2), 100495 (2024).
25. Zhang, S. et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Syst. Appl.* **237**, 121692 (2024).
26. Gladys, A. A. & Vetriselvi, V. Survey on multimodal approaches to emotion recognition. *Neurocomputing* **556**, 126693 (2023).
27. Kim, K. & Park, S. AOBERT: All-modalities-in-One BERT for multimodal sentiment analysis. *Inform. Fusion* **92**, 37–45 (2023).
28. Chen, C. & Aleem, M. A model for new media data mining and analysis in online english teaching using long short-term memory (LSTM) network. *PeerJ. Comput. Sci.* **10**, e1869 (2024).
29. Zhao, Q. et al. Cross-modal attention fusion network for RGB-D semantic segmentation. *Neurocomputing* **548**, 126389 (2023).
30. Zhao, L., Yang, Y. & Ning, T. A Three-stage multimodal emotion recognition network based on text low-rank fusion. *Multimedia Syst.* **30**(3), 142 (2024).
31. Yu, R. et al. *Weighted Multi-modal Contrastive Learning based Hybrid Network for Alzheimer's Disease Diagnosis* (IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2025).
32. Jang, H. et al. Modality-agnostic self-supervised learning with meta-learned masked auto-encoder. *Adv. Neural. Inf. Process. Syst.* **36**, 73879–73897 (2023).
33. Wang, L. et al. Automatic depression prediction via cross-modal attention-based multi-modal fusion in social networks. *Comput. Electr. Eng.* **118**, 109413 (2024).
34. Ezzameli, K. & Mahersia, H. Emotion recognition from unimodal to multimodal analysis: A review. *Inform. Fusion* **99**, 101847 (2023).
35. Zhao, Y. et al. Multimodal sentiment system and method based on CRNN-SVM. *Neural Comput. Appl.* **35**(35), 24713–24725 (2023).

Author contributions

Zhiming Liu: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation Tajularipin Sulaiman: writing—review and editing, visualization, supervision, project administration, funding acquisition Nur Raihan Che Nawi: methodology, software, validation, formal analysis.

Funding

This research received no external funding.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The studies involving human participants were reviewed and approved by Department of Curriculum and Instruction Studies, Faculty of Educational Studies, Universiti Putra Malaysia Ethics Committee (Approval Number: 2023.21540025). The participants provided their written informed consent to participate in this study. All methods were performed in accordance with relevant guidelines and regulations.

Additional information

Correspondence and requests for materials should be addressed to T.S.