

Multimodal Data-Driven Intelligent Assessment System for Classroom Teaching Quality—Based on Deep Learning Methods and Practices

Manyi Li¹, Donghe Liang², Wen Zuo²

¹Faculty of Humanities and Arts, Xi'an Fanyi University, Xi'an 710105, Shaanxi, China

²Faculty of Economics and Management, Xi'an Fanyi University, Xi'an 710105, Shaanxi, China

Corresponding author: Wen Zuo (e-mail: zouwen128182@163.com)

ABSTRACT With the rapid advancement of intelligent sensing technologies and Electromagnetic Waves, Antennas and Propagation systems, efficient multimodal information fusion and temporal perception have become fundamental requirements for communication-oriented monitoring and intelligent decision-making. To address the challenges of heterogeneous data integration and subjective classroom assessment, this study proposes a Hierarchical Cross-modal Attention with Temporal Dynamics Perception Network (HCA-TDPNet) for multimodal teaching quality evaluation. The framework integrates Video Swin Transformer, Wav2Vec 2.0, and BERT to extract visual, acoustic, and semantic representations, while a multi-head cross-attention mechanism and a hierarchical temporal encoder composed of temporal convolutional networks and gated recurrent units enable deep cross-modal interaction and long-term temporal dependency modeling. The resulting multimodal representations are mapped to quantitative indicators of teaching objective achievement, teacher-student interaction quality, and student participation depth. Experimental validation based on public pre-training datasets and the SCB-Dataset demonstrates Pearson correlation coefficients of 0.893, 0.918, and 0.904 for the three evaluation dimensions, respectively, while reducing the assessment time of a single lesson to approximately 2 minutes, saving more than 95% compared with manual evaluation. Beyond educational applications, the proposed architecture provides an effective framework for heterogeneous information fusion, real-time multimodal perception, and adaptive semantic interaction, offering valuable engineering references for intelligent sensing, communication-oriented monitoring, and distributed information processing in Electromagnetic Waves, Antennas and Propagation systems.

INDEX TERMS multimodal information fusion, cross-modal attention, temporal dynamic perception, intelligent sensing

I. INTRODUCTION

THE integration of deep learning and innovative educational technologies is reshaping contemporary teaching, learning, and assessment methods [1-2]. The deepening of digital transformation in education has led to the accumulation of massive multimodal data in the classroom teaching environment, including teaching videos, teacher-student voice interactions, blackboard texts, etc., providing an unprecedented data foundation for accurate evaluation of teaching quality. The evaluation of classroom teaching quality, as a core link to ensure and improve the level of education [3-4], is directly related to the scientific and objective development of teachers and students' academic achievements. The widely adopted expert evaluation mode is limited by labor costs and subjective experience, making

it difficult to achieve normalized and accurate monitoring of large-scale classrooms. This contradiction between practical needs and technological supply has given rise to a wave of research on intelligent evaluation systems.

The current construction of an intelligent evaluation system for classroom teaching quality faces multiple technical difficulties. The heterogeneity and semantic gap of multimodal data at the time scale make it difficult to effectively align and fuse different modal information such as visual, audio, and text [5-6]. Teacher actions, student expressions, and speech content in classroom scenes need to achieve spatiotemporal synchronization with millisecond level accuracy. Classroom teaching behavior has significant temporal dynamic characteristics [7], and a complete teacher-student interaction often spans several seconds to minutes. Tradi-

tional short-term feature extraction methods are difficult to capture this long-range dependency relationship [8]. The evaluation dimensions of teaching quality are highly abstract and cannot directly correspond to a single physical signal feature. It is necessary to establish a complex mapping from low-level perceptual data to high-level teaching semantics. The existing evaluation methods generally lack interpretability and are difficult to provide specific teaching behavior improvement suggestions to teachers.

In response to the above difficulties, the academic community has conducted a series of exploratory studies. Methods based on manual features and machine learning extract prosodic features from classroom audio or motion features from videos [9], and use classifiers such as support vector machines for classroom behavior recognition [10]. However, these methods are limited by the completeness of feature engineering and are difficult to generalize to complex and diverse real-world classroom scenarios. The method based on single-mode deep learning uses convolutional neural network to analyze teachers and students' attitudes in classroom images [11], or uses recurrent neural network to process voice transcribed text [12]. Although such methods have made progress in single-mode tasks, they ignore the semantic complementarity between different modal information, leading to blind spots in the perception of the overall atmosphere of the classroom. The multimodal fusion method based on simple feature concatenation directly concatenates the deep features of different modalities and inputs them into the classifier [13]. This method fails to solve the problem of temporal asynchrony between modalities, and linear concatenation cannot model the deep interaction relationships between modalities. Overall, existing methods still have significant shortcomings in multimodal deep semantic interaction and temporal dynamic perception, making it difficult to accurately quantify the connotation of classroom teaching quality.

To address the challenges of heterogeneous fusion of multimodal classroom data and insufficient dynamic perception in long-term teaching, this paper proposes HCA-TDPNet. This method first uses Video Swin Transformer to extract the spatiotemporal features of the spatial position and action posture of teachers and students in classroom video frames. The Wav2Vec 2.0 model is used for self supervised representation learning of classroom audio to capture intonation and interactive acoustic events in the language flow. The BERT model is introduced to map audio transcribed text into semantic embedding vectors. To solve the semantic alignment problem between visual and audio text modalities in the temporal dimension, a cross modal interaction module based on multi head cross attention mechanism is constructed, which enables bidirectional information transmission between visual feature sequences and audio text fusion feature sequences at each time step, generating enhanced multimodal representations containing deep teaching semantic interaction relationships. The representation is input into a hierarchical temporal

encoder consisting of a temporal convolutional network and gated loop units, which capture the evolution laws of short-term teaching behavior and long-term teaching processes, respectively. Finally, the classroom global representation is mapped to quantitative scores in dimensions such as teaching goal achievement, teacher-student interaction quality, and student participation depth through global attention pooling and multi head regression layers.

To verify the effectiveness of the model in evaluating the quality of real classroom teaching, this study adopts the strategy of "public pre-training+real classroom verification". The model first undergoes multi task pretraining on five publicly available multimodal datasets, namely DAiSEE, NoXi, CMU-MOSI, IEMOCAP, and MELD, to learn general abilities in emotion recognition, teacher-student interaction quality, and engagement representation. On this basis, SCB-Dataset is introduced as a real classroom validation dataset. By fine-tuning and evaluating on real classroom data, the three-dimensional teaching quality score output by the system is correlated and analyzed with the comprehensive behavior indicators extracted from teaching behavior annotations to verify the effectiveness of the model in real teaching scenarios. This article verifies the effectiveness and generalization ability of the model in actual teaching scenarios through fine-tuning and evaluation on real classroom data.

II. HCA-TDPNET: HIERARCHICAL CROSS MODAL ATTENTION AND TEMPORAL DYNAMIC PERCEPTION NET

OVERALL SYSTEM ARCHITECTURE

To achieve intelligent evaluation of classroom teaching quality, this article constructs the HCA-TDPNet system. Figure 1 presents the overall architecture of the HCA-TDPNet system, demonstrating the collaborative relationship and data flow of each module in the hierarchical architecture.

Figure 1 shows the overall architecture of the intelligent evaluation system for classroom teaching quality, consisting of four layers: system input, core model, system output, and deployment support. The input layer of the system receives classroom recordings, which are preprocessed and separated from audio and video. The visual feature sequence V and the audio text fusion feature sequence A are extracted separately. The core model consists of two parts: a cross modal attention fusion module and a hierarchical temporal encoder. In the cross modal attention fusion module, the visual feature sequence V and the audio text fusion feature sequence A are used as query and key value pairs, respectively. Two multi head cross attention units are used to achieve bidirectional semantic interaction between vision and audio text, and audio text to vision. Then, residual connections and layer normalization are used to merge them into an enhanced multimodal representation sequence H in the feature concatenation layer; The hierarchical temporal encoder inputs H into a Temporal Convolutional Network (TCN) to capture short-term teaching behavior patterns

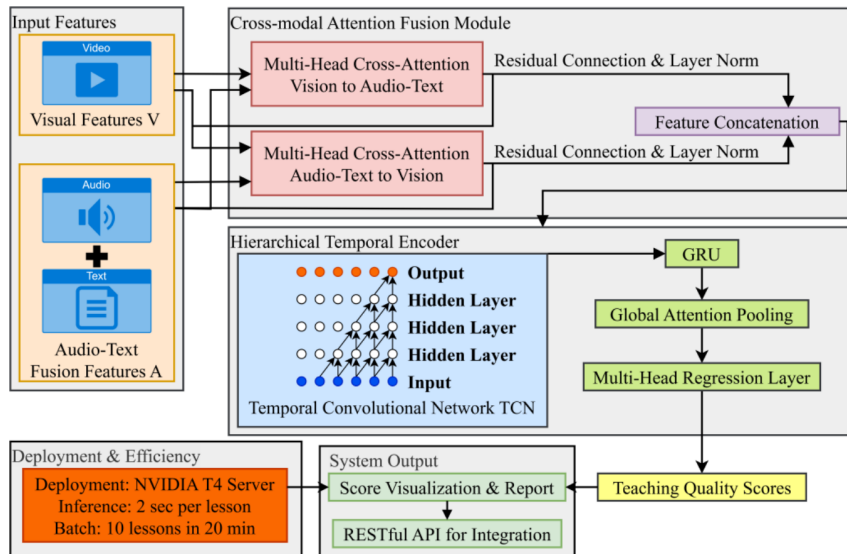


FIGURE 1. Overall architecture diagram of HCA-TDPNet

at the millisecond to second level. The Gated Recurrent Unit (GRU) models the evolution of long-term classroom segments at the second to minute level, and the global attention pooling layer focuses on key teaching segments. Finally, the multi head regression layer outputs quantitative scores in three dimensions: teaching objective achievement, quality of teacher-student interaction, and depth of student participation in parallel. The system output layer presents the scoring results through a visual interface and encapsulates them into a RESTful API (Representational State Transfer Application Programming Interface) to interface with the teaching management platform. The deployment support layer is based on NVIDIA T4 servers, with a single lesson inference time of 2 seconds and support for batch processing, providing engineering support for regular teaching monitoring. This architecture achieves semantic alignment at the time step level between modalities through cross modal attention, and combines hierarchical temporal modeling of TCN and GRU to complete end-to-end mapping from heterogeneous classroom perception data to multidimensional teaching quality ratings.

MULTIMODAL FEATURE EXTRACTION AND PREPROCESSING

The raw data required for evaluating the quality of classroom teaching includes video streams, audio streams, and transcribed texts from classroom recordings. Feature extraction and time alignment are performed on these three modalities of data.

The input is a sequence of consecutive frames $V = \{v_1, v_2, \dots, v_T\}$ of classroom images or videos, with each frame being an RGB image. The Video Swin Transformer model pre-trained on the large-scale action recognition dataset Kinetics-400 is used to extract spatiotemporal features. This model captures intra-frame spatial structure and inter-frame motion information through a shift window

attention mechanism [14], and outputs a feature tensor F_v with dimension $T \times H \times W \times C_v$, where T is the number of time steps, H and W are spatial feature map sizes, and C_v is the number of channels. The specific calculation can be expressed as:

$$F_v = \text{VideoSwin}(V; \theta_v) \quad (1)$$

In the formula, θ_v represents the pre-training parameters of Video Swin Transformer, which are partially unfrozen during the final stages of fine-tuning to allow the model to adapt to specific “classroom” visual semantics, such as teacher-student spatial positioning and blackboard interaction patterns, which significantly differ from the general actions in the Kinetics-400 dataset.

Separate the original audio signal A from the classroom video. Use the Wav2Vec 2.0 self supervised model [15-16] for acoustic feature extraction. This model is pre-trained on large-scale unlabeled speech data through contrastive learning and outputs frame level speech representations. Input A into the Wav2Vec 2.0 encoder to obtain an audio feature sequence $F_a = \{a_1, a_2, \dots, a_{T'}\}$, with each feature vector having a dimension of 768 and a temporal resolution related to the original audio sampling rate. To align with the video features in the time dimension, F_a is resampled [17-18] using linear interpolation to the same time step T as the video frame rate.

Audio A was transcribed into a text sentence sequence $S = \{s_1, s_2, \dots, s_K\}$ using an automatic speech recognition system, with each sentence corresponding to a discourse segment. The BERT-base [19-20] model was used to map each sentence into a semantic embedding vector, and the output labeled [CLS] was taken as the sentence-level representation to obtain the text feature sequence $F_t = \{t_1, t_2, \dots, t_K\}$, with each vector having a dimension of 768. Since discourse segments have start and end timestamps on the timeline, F_t was aligned to the video

timeline based on the timestamps to generate a text feature sequence F'_t synchronized with the video. Time periods without discourse were filled with a learnable “silence” embedding vector instead of simple zero vectors. This approach prevents “silent bias” and noise in the BERT embeddings, as the crossmodal attention mechanism can then explicitly distinguish between intentional instructional pauses and data gaps, thereby resolving potential temporal alignment ambiguities.

After completing the single-modal feature extraction, the visual features F_v , aligned audio features F'_a , and text features F'_t are concatenated at each time step to form a multimodal feature sequence $X = \{x_1, x_2, \dots, x_T\}$, where $x_t = [F_v(t), F'_a(t), F'_t(t)] \in \mathbb{R}^{C_v+768+768}$ is the modality. This sequence retains the original information of each modality while achieving precise temporal synchronization. Table 1 summarizes the input data, core model, and output dimensions of each modal feature extraction module. Visual features are extracted from the video frame sequence by the Video Swin Transformer, outputting a four-dimensional tensor. Audio features were extracted from the original audio using Wav2Vec 2.0, interpolated, and synchronized with the video. Text features were extracted from ASR (Automatic Speech Recognition) transcribed text using BERT, aligned based on timestamps. The time-aligned joint feature sequence provided a unified multimodal input representation for subsequent modules.

TABLE 1. Multimodal feature extraction and preprocessing module configuration

Modality	Feature Extraction Model	Input Data	Output Dimension
Video	Video Swin Transformer	Video Frame Sequence	Feature Tensor $T \times H \times W \times C_v$
Audio	Wav2Vec 2.0	Raw Audio Signal	Feature Sequence $T' \times 768$
Text	BERT-base	Transcribed Sentences	Feature Sequence $K \times 768$
Alignment	Linear Interpolation & Timestamp Mapping	Audio Text Features, Timestamps	Aligned Features $T \times 768$
Fusion	Feature Concatenation	Visual Aligned Features, Audio Aligned Text Features	Joint Features $T \times (C_v + 1536)$

Table 1 summarizes the configuration information of the multimodal feature extraction and preprocessing modules, corresponding to the feature extraction models, input data, output dimensions, and subsequent alignment and fusion methods for video, audio, text, and time-aligned joint and fused features. The video modality uses a Video Swin Transformer pre-trained on Kinetics-400 to extract a four-dimensional feature tensor containing spatiotemporal information from a continuous video frame sequence. The audio modality uses the Wav2Vec 2.0 model to extract frame-level acoustic feature sequences from the original audio signal, which are then linearly interpolated and synchronized with the video frame rate. The text modality uses BERT-

base to map the text sentences transcribed from automatic speech recognition into semantic embedding vectors, which are then aligned to the video timeline based on the utterance timestamps. Aligned audio and text features are concatenated with visual features at each time step to form a unified joint feature sequence. This table clarifies the source, processing method, and output format of each modality feature extraction, providing standardized input data specifications for subsequent cross-modal interaction and hierarchical temporal modeling.

CROSS-MODAL ATTENTION FUSION MECHANISM

Classroom teaching quality assessment requires the model to understand the deep connections between teacher and student actions, vocal intonation, and discourse semantics. The joint feature sequence X simply concatenates visual, audio, and text features at each time step, failing to achieve semantic interaction between modalities. Therefore, a cross-modal interaction module based on multi-head cross-attention is designed, enabling bidirectional information transmission between visual features and audio-text fusion features at each time step, generating an enhanced multimodal representation rich in teaching semantic relationships. The joint feature sequence X is split along the feature dimension into visual feature component $V \in \mathbb{R}^{T \times d_v}$ and audio-text fusion feature component $A \in \mathbb{R}^{T \times d_a}$, where $d_v = C_v$ and $d_a = 1536$. The cross-modal interaction module uses V as the query and A as the key and value to calculate the attention response of the visual modality to the audio-text modality. Simultaneously, using A as the query and V as the key and value, it calculates the attention response of the audio-text modality to the visual modality. Taking the visual-to-audio-text attention calculation as an example, a linear transformation maps the input to query $Q_v = VW_Q$, key $K_a = AW_K$, and value $V_a = AW_V$, where $W_Q \in \mathbb{R}^{d_v \times d_k}$, $W_K \in \mathbb{R}^{d_a \times d_k}$, and $W_V \in \mathbb{R}^{d_a \times d_v}$ are learnable parameter matrices, and d_k is the attention head dimension. The attention output is calculated as follows:

$$\text{Attn}_{v \rightarrow a} = \text{softmax}\left(\frac{Q_v K_a^T}{\sqrt{d_k}}\right) V_a \quad (2)$$

In the formula, $Q_v K_a^T$ generates an attention weight matrix with dimension $T \times T$, where each element represents the intensity of attention of visual features at a certain time step to audio-text features at other time steps. After softmax normalization [21], it is weighted and summed with V_a to obtain the visual enhancement feature $V' \in \mathbb{R}^{T \times d_v}$ that integrates audio-text semantics. Similarly, the attention output $A' \in \mathbb{R}^{T \times d_a}$ of the audio-text modality is calculated with A as the query and V as the key. The multi-head mechanism executes the above calculation in parallel h times, using different linear transformation parameters each time, and concatenates the outputs of h heads and then performs a linear transformation again to obtain the final outputs V_{mh} and A_{mh} . The features after cross-modal interaction are residually connected and layer normalized [22-

23] to obtain $V_{\text{final}} = \text{LayerNorm}(V + V_{\text{mh}})$ and $A_{\text{final}} = \text{LayerNorm}(A + A_{\text{mh}})$, which are then concatenated along the feature dimension to generate an enhanced multimodal feature sequence $H = [V_{\text{final}}, A_{\text{final}}] \in \mathbb{R}^{T \times (d_v + d_a)}$. The vector at each time step in this feature sequence contains both the visual information at the current moment and the audio-text semantic information after cross-modal interaction. The visual features carry the action response related to the speech content, and the audio-text features integrate the semantic representation matching the posture of teachers and students.

HIERARCHICAL TEMPORAL DYNAMIC MODELING

Classroom teaching processes exhibit complex multi-scale temporal characteristics. Teachers asking questions and students raising their hands last from hundreds of milliseconds to several seconds, while group discussions and lectures span several minutes. A hierarchical temporal encoder composed of a TCN and a gated recurrent unit (GRU) is constructed to capture dynamic patterns from short-term teaching behaviors to long-term classroom segments. Feature sequences $H \in \mathbb{R}^{T \times D}$ ($D = d_v + d_a$) are input into the TCN module. The TCN expands the receptive field by stacking multiple dilated convolutional layers, with each convolutional layer followed by weight normalization [24–25] and the ReLU (Rectified Linear Unit) activation function [26]. The dilated convolutional operation of the l -th layer is defined as:

$$Z_t^{(l)} = \sum_{k=0}^{K-1} W_k^{(l)} \cdot Z_{t-d^{(l)} \cdot k}^{(l-1)} \quad (3)$$

In the formula, K is the kernel size, $d^{(l)}$ is the dilation rate of the l -th layer, $Z^{(l-1)}$ is the output of the previous layer, $Z^{(0)} = H$, and $W_k^{(l)}$ are learnable kernels. The dilation rate increases exponentially with the number of layers, enabling the top-level neurons to cover a longer time range. The TCN module stacks L layers in total, with each layer maintaining the same output dimension as the input, and finally outputting $Z \in \mathbb{R}^{T \times D}$. This module extracts short-term behavioral features from milliseconds to seconds through local temporal convolution [27], such as the start and end of the teacher's gestures and the instantaneous changes in the student's facial expressions. Z is input into the GRU module for long-term dependency modeling. GRU controls the flow of information through update gates and reset gates [28–29], and the calculation at time step t is:

$$r_t = \sigma(W_r Z_t + U_r h_{t-1} + b_r) \quad (4)$$

$$u_t = \sigma(W_u Z_t + U_u h_{t-1} + b_u) \quad (5)$$

$$\tilde{h}_t = \tanh(W_h Z_t + U_h (r_t \odot h_{t-1}) + b_h) \quad (6)$$

$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot \tilde{h}_t \quad (7)$$

In formulas (4), (5), (6), and (7), Z_t is the input of the GRU at time t , h_{t-1} is the hidden state at the previous time step, r_t and u_t are the reset and update gates respectively, σ

is the sigmoid function [30–31], and $\odot$ represents element-wise multiplication. The hidden state $h_t \in \mathbb{R}^{D_h}$ of the GRU encodes the dynamic information of the classroom up to the current time step, and D_h is the dimension of the hidden layer. By stacking the hidden states of all time steps, the output $H_{\text{gru}} \in \mathbb{R}^{T \times D_h}$ is obtained. This sequence contains the evolutionary rules of teaching links from the second level to the minute level, such as the overall participation change in the discussion link and the logical coherence of the explanation link. In order to aggregate the variable-length temporal information into a fixed-dimensional global classroom representation, a global attention pooling layer is introduced [32–33]. This layer calculates the attention weight of the hidden state at each time step. The weight is determined by the correlation between the learnable query vector $q \in \mathbb{R}^{D_h}$ and the hidden state at each time step:

$$\alpha_t = \frac{\exp(q^T \tanh(W_\alpha H_{\text{gru}}[t] + b_\alpha))}{\sum_{j=1}^T \exp(q^T \tanh(W_\alpha H_{\text{gru}}[j] + b_\alpha))} \quad (8)$$

In the formula, $H_{\text{gru}}[t]$ represents the hidden state at time step t , $W_\alpha \in \mathbb{R}^{D_h \times D_h}$ and $b_\alpha \in \mathbb{R}^{D_h}$ are learnable parameters, and q is the learnable query vector. The final global classroom representation g is the weighted sum of the hidden states at all time steps:

$$g = \sum_{t=1}^T \alpha_t H_{\text{gru}}[t] \quad (9)$$

In the formula, $g \in \mathbb{R}^{D_h}$ integrates the key temporal information of the entire lesson, with the attention weight α_t automatically focusing on time periods that contribute significantly to teaching quality, such as peak teacher-student interaction or key explanation segments. To enhance the interpretability of the model output, the attention weight α_t output from the global attention pooling layer is explicitly retained and used for subsequent analysis. This weight sequence directly reflects the model's attention to each time step when generating the final score; a higher weight value indicates a greater contribution of the teaching behavior during that time period to the quality score. By performing correlation analysis between α_t and the teaching behaviors labeled in the SCB-Dataset, the mapping relationship between model scores and specific teaching behaviors can be revealed, thus providing teachers with actionable improvement suggestions such as “low quality of teacher-student interaction due to insufficient questioning time.” The global representation g is input into a multi-head regression layer [34–35] to predict scores. The mapping logic is grounded in educational psychology and pedagogical assessment theories: visual features capture behavioral engagement (postures, gestures), audio-text features capture cognitive engagement (question-answering logic, semantic coherence), and their fusion represents the holistic teaching-learning process. Specifically, the “teaching objective” score is derived from the alignment between the teacher's explained content (textual semantic embeddings) and the struc-

tural progression of the lesson (temporal dynamics). This layer consists of M independent fully connected networks, each corresponding to an evaluation index. The output of the m -th regression head is:

$$\hat{y}_m = W_m^{(2)} \text{ReLU}(W_m^{(1)} g + b_m^{(1)}) + b_m^{(2)} \quad (10)$$

In the formula, $W_m^{(1)} \in \mathbb{R}^{D_h \times D_h}$, $b_m^{(1)} \in \mathbb{R}^{D_h}$, $W_m^{(2)} \in \mathbb{R}^{1 \times D_h}$, and $b_m^{(2)} \in \mathbb{R}$ are the m -th learnable head parameters. The final output vector $\hat{y} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_M]$ is the system's evaluation score for the lesson across M quality dimensions. Table 2 summarizes the core parameter configurations of each module of the hierarchical temporal encoder. The TCN module uses 4 layers of dilated convolutions with dilation rates of 1, 2, 4, and 8, a kernel size of 3, and the output channel remains the same as the input dimension D . The hidden layer dimension D_h of the GRU module is set to 512. The query vector q and transformation matrix W_α in the global attention pooling layer are randomly initialized. The multi-head regression layer contains three independent heads, corresponding to the three evaluation dimensions: teaching objective achievement, quality of teacher-student interaction, and depth of student participation. Each head has a hidden layer dimension of 256. This configuration effectively controls computational complexity while maintaining model capacity.

TABLE 2. Parameter configuration of hierarchical temporal encoder

Module	Parameter Name	Parameter Value	Description
TCN	Number of Convolutional Layers	4	Expanding receptive field via stacked dilated convolutions
TCN	Dilation Rates	[1,2,4,8]	Exponential growth to cover multi-scale temporal patterns
TCN	Kernel Size	3	Temporal neighborhood size per convolutional layer
TCN	Output Channels	D	Consistent with input feature dimension
GRU	Hidden Dimension	512	Hidden state size for encoding long-term teaching dynamics
GRU	Number of Layers	1	Single-layer GRU to reduce overfitting risk
Global Attention	Query Vector Dimension	512	Learnable query vector dimension matching hidden state
Global Attention	Transformation Matrix Dimension	512×512	Projection matrix for attention score computation
Multi-Head Regression	Number of Heads	3	Corresponding to three teaching quality assessment dimensions
Multi-Head Regression	Hidden Dimension	256	Hidden layer dimension within each regression head

Table 2 shows the core parameter configurations of each module in the hierarchical temporal encoder, covering four key components: TCN, GRU, global attention pooling, and

multi head regression layer. The TCN module adopts four layers of dilated convolutions with dilation rates of 1, 2, 4, and 8, and a convolution kernel size of 3. The output channel is consistent with the input feature dimension D . The coverage of multi-scale temporal patterns is achieved through exponentially increasing dilation rates. The hidden layer dimension of the GRU module is set to 512, with a layer size of 1, used to encode long-term teaching dynamics ranging from seconds to minutes. The query vector dimension and transformation matrix dimension in the global attention pooling layer are both 512, focusing on key teaching segments through learnable attention weights. The multi head regression layer consists of three independent heads, corresponding to three output dimensions: teaching objective achievement, quality of teacher-student interaction, and depth of student participation. The hidden layer dimension within each head is 256. This table specifies the hyperparameter settings for each module during the time series modeling process, ensuring the reproducibility of the model architecture.

SYSTEM DEPLOYMENT AND INFERENCE EFFICIENCY

To achieve the large-scale application of the classroom teaching quality evaluation system in real teaching scenarios, explain the deployment method and inference efficiency of the system from an engineering perspective. The model is trained and inferred based on the PyTorch 1.12 framework. After offline training and validation, the model is converted and optimized using ONNX (Open Neural Network Exchange) format and deployed to NVIDIA T4 GPU servers. The system adopts a server-side deployment model, supporting

both batch processing and real-time inference. It can interface with existing teaching management platforms via a RESTful API (Representational State Transfer Application Programming Interface) to receive classroom recordings and return teaching quality assessment results. In terms of inference efficiency, a single inference iteration takes 2.0 seconds, and the complete evaluation process for a single lesson is controlled within 2 minutes. The system uses a batch processing mode, allowing parallel processing of multiple classroom recordings. Benefiting from the parallel architecture of the T4 GPU, processing 10 lessons simultaneously takes a total of only 8 minutes, achieving a significant speedup compared to the 20 minutes required for sequential processing. The model has 128M parameters, and the memory usage is approximately 2.5GB when using FP16 half-precision inference, meeting the deployment constraints of conventional teaching resource platforms. To adapt to edge computing or resource-constrained scenarios, the system reserves interfaces for knowledge distillation and quantization compression. The aforementioned deployment scheme and efficiency indicators provide engineering feasibility support for the system's routine monitoring application in actual teaching environments.

III. EXPERIMENTAL DESIGN AND DATA DESCRIPTION

EXPERIMENTAL DATASET

This study adopts a data strategy of “public pre-training + real classroom validation,” including two types of datasets. The public multimodal dataset is used for basic model capability learning, while the real classroom dataset is used for model fine-tuning and final evaluation.

PUBLIC PRE-TRAINING DATASET

The model undergoes multi task pre-training on the following five publicly available multimodal datasets to learn general abilities such as emotion recognition and teacher-student interaction quality:

Table 3 summarizes the information of five publicly available multimodal datasets used for model pre-training, including dataset name, modality type, sample size, original annotation content, and their mapping tasks in this paper. The DAiSEE dataset contains 9068 facial video clips of students, annotated with emotional states such as engagement and boredom, for student engagement assessment. The NoXi dataset includes 30 hours of expert discussion conversations, including video, audio, and transcribed text, annotated with interest levels and interaction quality, for modeling teacher-student interaction quality. The CMU-MOSI dataset provides 2199 movie review clips, annotated with emotional intensity, while cinematic expressions are more hyperbolic than classroom interactions, this dataset is utilized solely to assist the model in learning high-variance affective features and extreme acoustic-semantic mappings, which enhances the network’s sensitivity to subtle emotional fluctuations in professional settings during the fine-tuning stage, for extreme classroom atmosphere assisted learning. The IEMOCAP dataset includes 12 hours of two person conversations, labeled with emotion categories such as happiness and sadness, for identifying emotional changes between teachers and students. The MELD dataset contains over 1400 multi-turn dialogue videos, labeled with emotions and sentiments, for temporal modeling in complex interaction scenarios. Table 3 illustrates the size, modal composition, and mapping application of each pre-training dataset in this paper’s task, providing a basis for the rationale of joint training across datasets.

REAL CLASSROOM VALIDATION DATASET: SCB-DATASET

To validate the model’s effectiveness in assessing the quality of real classroom teaching, this study uses the SCB-Dataset (Student Classroom Behavior Dataset), publicly available at <https://github.com/Whiffe/SCBdataset>. This dataset contains 13,330 images and 122,977 labels, comprehensively recording teacher lectures, teacher-student interactions, and student activities. This dataset contains annotations of classroom teaching behaviors and calculates the duration proportions of six types of teaching behaviors (teacher questioning, student answering, teacher explanation, student discussion, blackboard demonstration, classroom practice) in the dataset. The

classroom samples are divided into three categories: lecture dominant, interactive dominant, and practice dominant, which are used for evaluating the model’s generalization ability. The SCB-

Dataset images selected in this article are divided into training set, validation set, and test set in an 8:1:1 ratio. To construct an objective verification benchmark, this article calculates the normalized duration ratio of each behavior category in each class based on the annotations of six teaching behaviors in the dataset. On this basis, a comprehensive behavioral indicator is constructed by weighting expert experience. Specifically, following the ‘Delphi Method’ and consultations with senior pedagogical experts, weights are assigned as 0.4 for teacher questioning and answering (recognizing its core role in active learning), 0.3 for teacher explanation and blackboard writing, and 0.3 for student discussion and practice. This provides an objective reference for correlation analysis with system output. After pre-training is completed, the model is fine tuned on the training set and the Pearson correlation coefficient between the three-dimensional score output by the system and the comprehensive behavioral indicators is calculated on the test set.

EXPERIMENTAL SETUP

The model is implemented based on the PyTorch 1.12 framework and trained and inferred on a single NVIDIA A100 80GB GPU (Graphics Processing Unit). Video feature extraction uses a Video Swin Transformer Tiny version pre-trained on Kinetics-400, audio feature extraction uses the Wav2Vec 2.0 Base model, and text feature extraction uses BERT-base-uncased. The parameters of the three pre-trained models are frozen during training. The cross-modal attention module has 8 attention heads with a dimension of 64. The TCN module contains 4 dilated convolutional layers with dilation rates of 1, 2, 4, and 8, and a kernel size of 3. The output channels of each layer maintain the same dimension as the input. The GRU module has a hidden layer dimension of 512 and a layer count of 1. The dimensions of the query vector and transformation matrix in the global attention pooling layer are both 512. The multi head regression layer consists of three independent heads, corresponding to three output dimensions: teaching objective achievement, quality of teacher-student interaction, and depth of student participation. Each head has an internal hidden layer dimension of 256. The training uses the AdamW optimizer with an initial learning rate of $1e-4$ and a weight decay of 0.01. Set the batch size to 16 and the number of training rounds to 50. The loss function is the mean square error, which calculates the error between the three-dimensional score vector output by the model and the normalized value of the teaching score. During the training process, an early stop strategy is used to stop training when the validation set loss does not decrease for 5 consecutive rounds. During data loading, the video is randomly cropped to a resolution of 224×224 , the audio is resampled to 16kHz, and the text is truncated to a

TABLE 3. Information of pre-trained multimodal dataset

Dataset	Modality	Sample Size	Original Annotation	Mapped Task
DAiSEE	Video	9068 clips	Engagement, Boredom, Confusion, Frustration	Student Engagement
NoXi	Video+Audio+Text	30 hours	Interest Level, Interaction Quality	Teacher-Student Interaction Quality
CMU-MOSI	Video+Audio+Text	2199 clips	Sentiment Intensity	Classroom Atmosphere Positivity
IEMOCAP	Video+Audio+Text	12 hours	Emotion Categories	Teacher-Student Emotion Variation
MELD	Video+Audio+Text	1400+ dialogues	Emotion and Sentiment	Complex Interaction Temporal Modeling

maximum BERT length of 512. All input features undergo layer normalization before being fed into the network.

COMPARISON METHODS AND EVALUATION INDICATORS

The comparison method includes two types of baselines, with the visual unimodal baseline using only the features extracted by the Video Swin Transformer and directly regressing after global average pooling. Feature concatenation+Long Short-Term Memory (LSTM) baseline concatenates three modal features in time steps and inputs them into two layers of LSTM for regression. The experiment uses Pearson correlation coefficient to evaluate the performance of the model. The Pearson correlation coefficient measures the linear correlation between the three-dimensional score vector output by the model and the comprehensive behavioral indicators. The calculation formula is:

$$\rho = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}} \quad (11)$$

In formula (11), n is the number of samples in the test set, y_i is the normalized value of the comprehensive behavioral indicator, $\hat{y}_i$ is the model prediction score, and $\bar{y}$ and $\bar{\hat{y}}$ are the corresponding means, respectively.

The value range of this indicator is $[-1, 1]$, and the closer it is to 1, the stronger the consistency between the model output and the trend of the comprehensive behavior indicator. Accuracy is used to evaluate the classification performance of the model in student engagement assessment tasks. The model's predicted student engagement scores are thresholded into three levels: high, medium, and low, and compared with the engagement level annotations of the dataset to calculate the proportion of correctly predicted samples in the total sample.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

CORRELATION ANALYSIS BETWEEN SYSTEM SCORING AND TEACHING BEHAVIOR INDICATORS

To evaluate the intrinsic correlation between system scoring and objective classroom behavior, a comparative analysis was conducted between the system's outputs in three dimensions: teaching objective achievement, quality of teacher-student interaction, and depth of student participation, and comprehensive behavioral indicators extracted from SCB-Dataset annotations. The comprehensive behavior index is based on the normalized duration proportion of six types of teaching behaviors, and is constructed by weighting

expert experience weights to reflect the overall activity and structural characteristics of teaching behaviors in the classroom. 30 test samples were selected for the experiment, and each sample was generated by the system with three

dimensions of predicted scores. At the same time, corresponding comprehensive behavioral indicators were extracted from the annotations to form 30 sets of paired data.

Figure 2 (a) corresponds to the teaching objective achievement, and the blue scatter plot shows a clear positive correlation trend. The Pearson correlation coefficient in the upper left corner is $\rho=0.893$, indicating that the system's judgment of the teaching objective achievement is highly consistent with behavioral indicators such as the duration of teacher lectures and the frequency of blackboard demonstrations in the classroom. Figure 2 (b) corresponds to the quality of teacher-student interaction, and the scatter plot also shows a good linear trend with a correlation coefficient of $\rho=0.918$. Some high interaction samples correspond to intensive periods of teacher questioning and student answering behavior, verifying the system's accurate capture of interactive behavior. Figure 2 (c) corresponds to the depth of student participation, with a correlation coefficient of $\rho=0.904$. The scatter points are symmetrically distributed on both sides of the diagonal, and the high participation samples correspond to classroom segments with higher student discussion duration and participation rate in classroom exercises.

The above results are attributed to the design mechanism of HCA-TDPNet proposed in this paper. The degree of teaching objective achievement is highly correlated with behavioral indicators such as the duration of teacher lectures and the frequency of blackboard demonstrations. This is due to the model capturing the logic of lectures through text modality and recognizing the rhythm of blackboard writing through visual modality; The quality of teacher-student interaction is strongly correlated with the intensive periods of questioning and answering behavior, stemming from the model's recognition of question and answer rounds through audio modality and capturing teacher-student posture turns through visual modality; The depth of student participation is strongly correlated with the duration of student discussions and classroom exercises, stemming from the model's comprehensive judgment of active participation behavior through visual and audio modalities. The hierarchical temporal encoder captures second level behavior patterns through temporal convolutional networks, and then models minute level teaching process evolution using gated loop

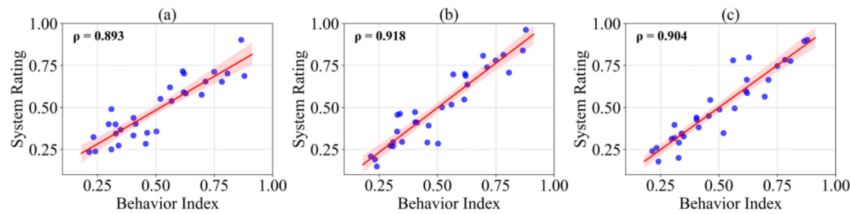


FIGURE 2. Correlation between system scoring and teaching behavior indicators: (a) The correlation between the teaching objective achievement and teaching behavior indicators; (b) The correlation between the quality of teacher-student interaction and teaching behavior indicators; (c) The correlation between student participation depth and teaching behavior indicators

units, making the system particularly sensitive to dynamic indicators such as “student engagement depth”, with the highest correlation coefficient. The correlation of 0.893 or above in all three dimensions indicates that cross modal information complementarity and multi-scale temporal modeling effectively solve the problem of incomplete recognition of teaching behavior and insufficient modeling of temporal dependencies in traditional single modal methods, achieving end-to-end mapping from heterogeneous classroom data to high-quality evaluation.

GAIN OF MULTIMODAL FUSION STRATEGY ON PERFORMANCE EVALUATION

To investigate the impact of the fusion weights of visual, audio, and text modalities on the accuracy of classroom teaching quality evaluation, the experiment used a ternary phase diagram to traverse and analyze the weight space. Under the constraint that the sum of visual weight, audio weight, and text weight is 1, discretize the weight combination of the three modalities into uniform grid points within a triangle, where each point represents a fusion strategy, and calculate the accuracy of the model in the student participation evaluation task under this strategy. This article visualizes the continuous distribution of accuracy with weight combinations using contour filling method, and marks the optimal fusion point within the triangle to obtain the results shown in Figure 3.

The three vertices in Figure 3 correspond to visual weight, audio weight, and text weight, respectively. Any point within the triangle represents the fusion weight combination of the three modalities. As shown in Figure 3, the accuracy distribution in the ternary phase diagram presents a unimodal shape, with peak areas located near visual weight 0.5, audio weight 0.3, and text weight 0.2. The accuracy reaches 0.91, and gradually decreases to below 0.75 towards the three vertices. This distribution originates from the complementary characteristics of multimodal information. The visual modality provides spatial position and action posture information for teachers and students, the audio modality captures intonation and interactive rhythm in the language flow, and the text modality carries the semantic logic of teaching content. When the three modalities are integrated, there are differences in the marginal benefits contributed by each modality. The peak value tends to have higher visual and audio weights, reflecting that the contribution of visual and audio information to teaching

quality evaluation in classroom scenes is higher than that of pure text information. When the text weight is too high, the accuracy decreases due to the loss of speech rhythm and visual dynamic information in pure text transcription. When the weight of visual or audio single modality is too high, the accuracy also decreases due to the lack of semantic supplementation from other modalities. This verifies that the gain of multimodal fusion in evaluating performance comes from the collaborative complementarity of heterogeneous information.

VISUALIZATION OF TEMPORAL ATTENTION FOCUS IN TEACHING BEHAVIOR

To visualize the temporal dynamic perception differences of different models on classroom teaching behavior and demonstrate the attention focusing ability of the models on classroom timing, this section simulates a 50 minute classroom timing based on public classroom video data and generates attention weight distributions. The experiment extracted the attention weight distribution of six teaching behaviors, including teacher questioning, student answering, teacher explanation, student discussion, blackboard demonstration, and classroom practice, over a 50 minute classroom time sequence. This article calculates the attention weights corresponding to each behavior category at each time point for the HCA-TDPNet model, visual unimodal baseline model, and feature concatenation+LSTM baseline model. Three heatmaps are generated to compare the time accuracy and behavior discrimination of the three models in focusing on teaching behavior. The results shown in Figure 4 are obtained.

Figure 4 presents the temporal attention distribution of the HCA-TDPNet model and two baseline models in six types of teaching behaviors. The horizontal axis represents classroom time, and the vertical axis represents six types of behaviors: teacher questioning, student answering, teacher explanation, student discussion, blackboard demonstration, and classroom practice. Color brightness represents attention weights. In Figure 4 (a) of the model presented in this article, teacher questioning and blackboard demonstration form clear highlighted bands during the 5-15 minute period, student discussion focuses during the 30-40 minute period, teacher explanation focuses during the 20-30 minute period, and classroom practice peaks during the 45-50 minute period. The highlighted periods of each behavior coincide with the natural occurrence time of the teaching process, and

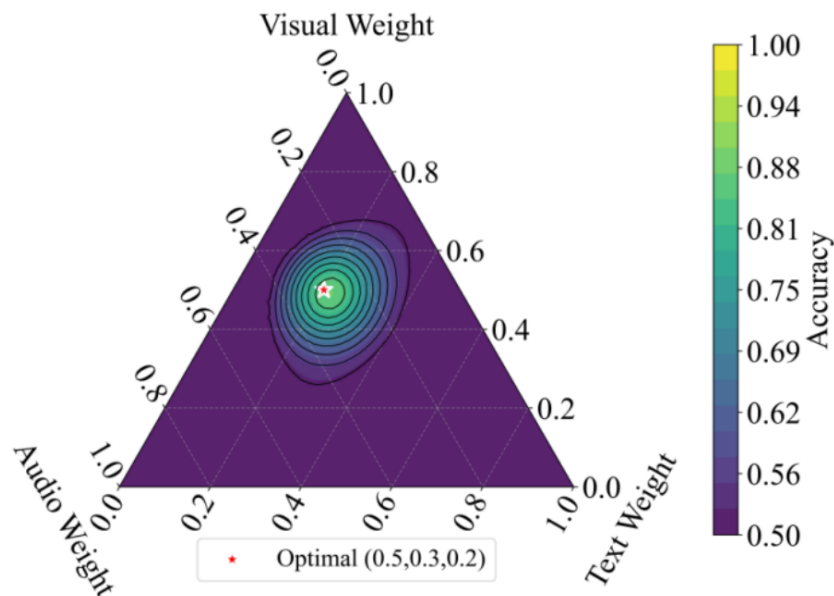


FIGURE 3. The influence of multimodal fusion weights on evaluation accuracy

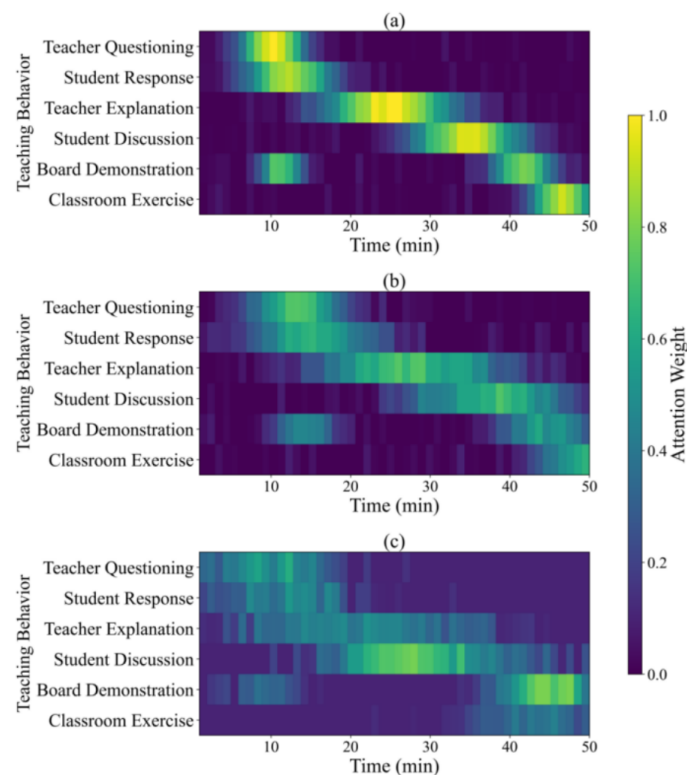


FIGURE 4. Comparison of temporal attention heatmaps for teaching behaviors of different models: (a) This article presents a heatmap of temporal attention for model teaching behavior; (b) Visual unimodal baseline teaching behavior temporal attention heatmap; (c) Splicing+LSTM baseline teaching behavior temporal attention heatmap

there is no significant overlap between different behaviors. Figure 4 (b) shows that the attention peak of the visual unimodal baseline has shifted, the peak shape has widened, and there is an inappropriate highlight during the period of student answers and teacher explanations. The highlighted area in Figure 4 (c) of the stitching+LSTM baseline ap-

pears more blurred, with false peaks overlapping during the teacher's explanation and student discussion periods, and inappropriate highlighting during the practice period for the blackboard demonstration behavior. The reason why the model in this article can achieve precise temporal focusing is due to its cross modal attention fusion mechanism, which

enables bidirectional information exchange between visual, audio, and textual features at every time step. This enables the model to accurately locate key teaching events based on the synchronization of teacher speech content and blackboard actions, as well as the temporal correlation between student answers and questions. The temporal convolutional network in the hierarchical temporal encoder captures short-term behaviors at the millisecond to second level, while the gated loop unit models the evolution of teaching processes at the second to minute level. The two work together to strictly limit attention weights to the natural occurrence periods of each behavior. The visual unimodal baseline relies solely on image sequences and lacks semantic temporal anchors provided by audio and text, resulting in its peak position deviating from the actual time of behavior occurrence and broadening of peak shape. Although concatenation+LSTM baseline introduces multimodality, feature concatenation cannot model deep interactions between modalities. Although LSTM can capture temporal dependencies, it cannot distinguish causal relationships between different behaviors, resulting in confusion and overlap of attention between multiple behaviors. The comparative results confirm that the combined effect of cross modal attention and hierarchical temporal modeling enables the model to accurately perceive the dynamic evolution of various teaching behaviors in the classroom.

EVALUATION OF MODEL INTERDISCIPLINARY GENERALIZATION ABILITY

Based on the differences in teaching behavior labeling in SCB-Dataset, classroom samples are divided into three categories according to the dominant type of teaching process: lecture dominant type (teacher explanation frequency $\geq 50\%$), interactive dominant type (teacher questioning+student answering frequency $\geq 40\%$), and practice dominant type (classroom practice frequency $\geq 30\%$). This article calculates the absolute errors between the systematic and manual ratings of HCA-TDPNet model, visual unimodal baseline model, and feature concatenation+LSTM baseline model in various types of classrooms. The error distribution characteristics of the three models in different classroom types are compared using grouped box plots, and the results shown in Figure 5 are obtained.

Figure 5 shows the absolute error distribution of three models in three types of classrooms: lecture led, interactive led, and practice led. The median error of the HCA-TDPNet model in this article is 0.062 in lecture led classrooms, 0.070 in interactive led classrooms, and 0.071 in practice led classrooms, with similar distribution concentrations. The median errors of visual unimodal baseline in three types of classrooms are 0.111, 0.120, and 0.117, respectively. The median errors of feature concatenation and LSTM baseline in three types of classrooms are 0.088, 0.097, and 0.084, respectively. The difference in distribution is due to the deep semantic interaction of visual, audio, and text features achieved by the model in this article through cross modal

attention mechanism, which enables the model to capture common teaching behavior patterns in different classroom types rather than surface features of specific teaching processes. In a lecture led classroom, the model captures the teacher's explanation logic through text modality and identifies the rhythm of the blackboard through visual modality to maintain low error. In interactive classrooms, the model captures question and answer cycles and speech overlap through audio modality, and identifies teacher-student posture changes through visual modality, effectively responding to high-frequency interactions. In a practice led classroom, the model recognizes students' exercise movements through visual modalities and reads exercise instructions through text modalities to maintain stable performance. The visual unimodal baseline lacks audio and text information, and overly relies on visual scene layouts with significant differences in various types of classrooms, resulting in increased errors and significant fluctuations between types. Although feature concatenation+LSTM baseline integrates multimodal information, it lacks semantic alignment on the timeline and has weaker generalization ability for different classroom types compared to the model proposed in this paper. The above results verify that the model in this article has lower errors and more concentrated distributions in different classroom scenarios, and has stronger generalization stability.

COMPARISON OF EFFICIENCY BETWEEN INTELLIGENT EVALUATION AND MANUAL EVALUATION

To quantify the time efficiency differences between intelligent evaluation systems and manual evaluations at different task scales, experiments were conducted to record the cumulative time spent on processing 1 to 10 classes for both manual and system evaluations. The manual evaluation time is based on the average time taken by three experimental personnel to independently complete classroom listening and evaluation classes, with each class lasting 45 minutes and an error range of ± 3 minutes per class; The system evaluation process includes the entire process of data pre-processing, model inference, and report generation, with 2 minutes per class and an error range of ± 0.2 minutes per class. Figure 6 shows the time consumption distribution and error range of two evaluation methods as the number of classes increases.

Figure 6 shows a comparison of the time spent on processing 1 to 10 class hours between manual evaluation and intelligent system evaluation. The horizontal axis represents time, the blue bar extending to the left represents the time spent on manual evaluation, the red bar extending to the right represents the time spent on system evaluation, and the vertical axis represents the number of class hours. The manual evaluation of a single class takes 45 minutes, and the processing of 10 classes accumulates to 450 minutes. The system evaluation takes 2 minutes for a single class, and processing 10 classes accumulates to 20 minutes. The difference in time consumption between the two evaluation methods is significant, and as the num-

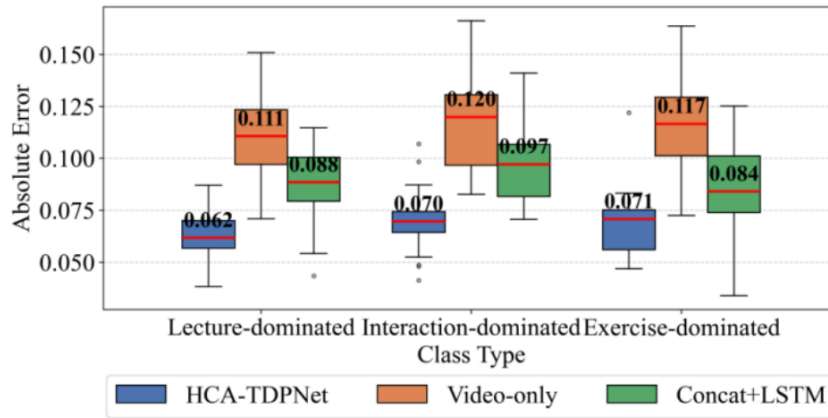


FIGURE 5. Comparison of absolute error distribution of different models in different classroom types

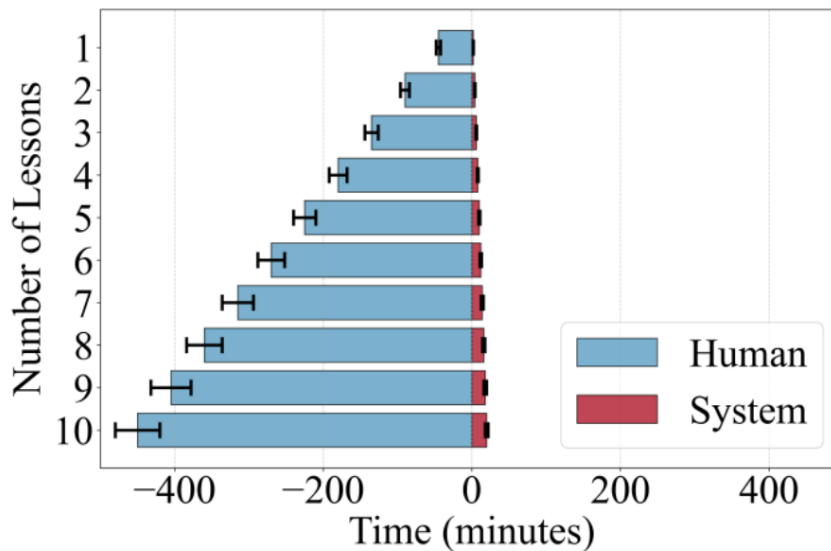


FIGURE 6. Comparison of efficiency between manual evaluation and system evaluation

ber of classes increases, the time consumption of manual evaluation increases linearly, while systematic evaluation maintains stable and high efficiency.

The efficiency difference stems from the fundamental difference in the evaluation mechanism. Manual evaluation relies on relevant personnel to listen, record, discuss, and grade the entire course. Each class requires a complete 45 minute real-time participation, and the evaluation process is synchronized with the teaching process, making it impossible to process multiple classes in parallel. The system evaluation is based on a pre-trained HCA-TDPNet model, which takes classroom videos as input and achieves end-to-end automated analysis through cross modal attention mechanism and hierarchical temporal encoder. The inference time for a single class is only 2 minutes, including data loading, feature extraction, and score generation. It can also batch process multiple classes, and the time cost is strictly linearly related to the number of classes. The difference in error lines further reflects the difference in stability between the two: manual evaluation is affected by the fatigue of

experimental personnel, subjective judgment fluctuations, and environmental factors, and errors accumulate as the evaluation amount increases; The system evaluation is based on a deterministic calculation process, and the error only comes from small fluctuations in the input data, resulting in a small cumulative error. This result validates the absolute efficiency advantage of intelligent evaluation systems over manual evaluation in large-scale, normalized classroom monitoring scenarios.

SYSTEM OUTPUT INTERPRETABILITY ANALYSIS

This article adopts the teaching behavior score correlation analysis as the core interpretability method, and uses the six types of teaching behavior frequency ratio annotation in the SCB-Dataset to display the score distribution of different models in six dimensions of teaching quality, and explains the formation mechanism of the model score in this article by combining the frequency ratio. To evaluate the scoring distribution characteristics of different models in terms of multidimensional teaching quality, six dimen-

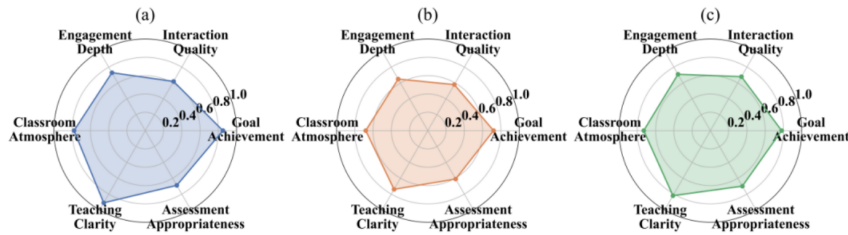


FIGURE 7. Radar chart of teaching quality dimension scores for different models: (a) HCA-TDPNet teaching quality dimension score radar chart; (b) Radar chart of visual unimodal baseline teaching quality dimension score; (c) Splicing+LSTM baseline teaching quality dimension score radar chart

sions were selected for the experiment: teaching objective achievement, quality of teacher-student interaction, depth of student participation, classroom atmosphere, clarity of teaching, and appropriateness of evaluation. The dimensional scores of the HCA-TDPNet model, visual unimodal baseline model, and feature concatenation+LSTM baseline model in the classroom samples were calculated separately, and the results shown in Figure 7 were obtained. Figure 7 presents the quantitative score distribution of the HCA-TDPNet model, visual unimodal baseline model, and feature concatenation+LSTM baseline model in six dimensions of teaching quality in this article. Figure 7 (a) shows that the model in this article has a score of 0.85 for achieving teaching objectives, 0.62 for teacher-student interaction quality, 0.73 for student engagement depth, 0.78 for classroom atmosphere, 0.91 for teaching clarity, and 0.69 for evaluation appropriateness, presenting clear inter dimensional differentiation. Figure 7 (b) shows that the visual unimodal baseline scores are concentrated in the range of 0.58 to 0.74 in each dimension, while Figure 7 (c) shows that the feature concatenation+LSTM baseline scores are distributed in the range of 0.68 to 0.82. The discrimination between the two dimensions is weaker than that of proposed model.

The score distribution of the model in this article reflects the underlying mechanism of cross modal attention fusion and hierarchical temporal encoding. The score of 0.62 for the quality of teacher-student interaction is lower than other dimensions, due to the model capturing pauses and overlaps in speech interaction through audio modality, and recognizing synchronous differences in teacher-student posture through visual modality. Overall, it is judged that the depth of interaction is insufficient. The highest score for teaching clarity is 0.91, as the logical structure explained by the teacher in the text mode is highly semantically aligned with the rhythm presented on the blackboard in the video mode. The visual unimodal baseline lacks audio and text information, making it difficult to perceive interactive rhythm and semantic logic, resulting in average and overall low scores in various dimensions. Although feature concatenation+LSTM baseline integrates multimodal features, the lack of semantic

alignment on the timeline results in temporal misalignment between visual and audio text features, leading to significant deviations in dimension scores that rely on cross modal synchronization judgments such as interaction quality and participation depth. The model in this article uses a multi head cross attention mechanism to achieve information exchange at the time step level between modalities, accurately mapping the scores of each dimension to the teaching reality and presenting a differentiated distribution consistent with artificial teaching evaluation cognition.

V. CONCLUSION

This article constructs HCA-TDPNet to address the difficulties in heterogeneous fusion of multimodal data and insufficient perception of long-term teaching dynamics in classroom teaching quality evaluation. After multi task pre-training on five publicly available multimodal datasets, the model was fine tuned and validated on the SCB-Dataset real classroom dataset. The experimental results showed that the Pearson correlation coefficient between the system and comprehensive behavioral indicators on the SCB-Dataset test set reached 0.893, 0.918, and 0.904 in the dimensions of teaching goal achievement, teacher-student interaction quality, and student participation depth, respectively. The single lesson evaluation was only 2 minutes long, saving more than 95% of time compared to manual evaluation. This study introduces a deep learning driven multimodal fusion mechanism into the field of classroom teaching evaluation, achieving end-to-end mapping from heterogeneous classroom data to multidimensional quality scoring. Its effectiveness and transferability in practical scenarios have been verified through real classroom data, providing scalable technical support for normalized teaching monitoring, accurate feedback on teacher professional development, and teaching improvement in the context of educational digital transformation.

AUTHOR CONTRIBUTIONS

Conceptualization – Manyi Li, Donghe Liang and Wen Zou; methodology – Manyi Li, Donghe Liang and Wen Zou; investigation – Manyi Li, Donghe Liang and Wen Zou; writing-original draft preparation – Manyi Li, Donghe Liang and Wen Zou. All authors have read and agreed to the published version of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflict of interest.

FUNDING

This research received no external funding.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] M. S. Far, "Artificial intelligence, innovative educational technologies, and transformation in classroom assessment," *Metaversalize*, vol. 2, no. 2, pp. 122-131, 2025, doi: 10.22105/metaverse.v2i2.70.
- [2] M. Kovacheva, "From Big Data To Innovation: Examining The Impact Of Digitalization And Artificial Intelligence On Education," *Narodnoopanski arxiv*, vol. (1), pp. 94-110, 2025, doi: 10.58861/tae.ea-nas.a.2025.1.05.en.
- [3] N. G. Bepari, R. Barua, and F. Rabbi, "Quality Assurance In Education: Innovative Approaches For Effective Classroom Management And Student Engagement," *Journal Of Creative Writing*, vol. 8, no. 3, pp. 1-22, 2024, doi: 10.1177/23476311221143231.
- [4] Y. Gao, "Deep learning-based strategies for evaluating and enhancing university teaching quality," *Computers and Education: Artificial Intelligence*, vol. 8, no. 1, Art. no. 100362, 2025, doi: 10.1016/j.caeai.2025.100362.
- [5] A. Barua, M. U. Ahmed, and S. Begum, "A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions," *Ieee access*, vol. 11, no. 1, pp. 14804-14831, 2023, doi: 10.1109/ACCESS.2023.3243854.
- [6] S. Song, X. Li, S. Li, et al., "How to bridge the gap between modalities: Survey on multimodal large language model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 9, pp. 5311-5329, 2025, doi: 10.1109/TKDE.2025.3527978.
- [7] W. Zhang, S. Gan, H. He, et al., "Effects of Teaching Behaviors on the Effectiveness of Classroom Learning Time," *International Journal of Emerging Technologies in Learning*, vol. 17, no. 9, pp. 214-227, 2022, doi: 10.3991/ijet.v17i09.30943.
- [8] M. Krichen and A. Mihoub, "Long short-term memory networks: A comprehensive survey," *AI*, vol. 6, no. 9, pp. 215, 2025, doi: 10.3390/ai6090215.
- [9] C. H. Heng, M. Toyoura, C. S. Leow, et al., "Analysis of Classroom Processes Based on Deep Learning With Video and Audio Features," *IEEE Access*, vol. 12, no. 1, pp. 110705-110712, 2024, doi: 10.1109/ACCESS.2024.3434742.
- [10] N. I. Mohd Talib, N. A. Abd Majid, and S. Sahran, "Identification of student behavioral patterns in higher education using Kmeans clustering and support vector machine," *Applied Sciences*, vol. 13, no. 5, pp. 3267, 2023, doi: 10.3390/app13053267.
- [11] T. Xu, W. Deng, S. Zhang, et al., "Research on recognition and analysis of teacher-student behavior based on a blended synchronous classroom," *Applied Sciences*, vol. 13, no. 6, pp. 3432, 2023, doi: 10.3390/app13063432.
- [12] Y. Kumar, A. Koul, and C. Singh, "A deep learning approaches in text-to-speech system: a systematic review and recent research perspective," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15171-15197, 2023, doi: 10.1007/s11042-022-13943-4.
- [13] F. Zhao, C. Zhang, and B. Geng, "Deep multimodal data fusion," *ACM computing surveys*, vol. 56, no. 9, pp. 1-36, 2024, doi: 10.1145/3649447.
- [14] T. Gao, Q. Zhang, T. Chen, et al., "Spatial-temporal sequence attention based efficient transformer for video snow removal," *Big Data Mining and Analytics*, vol. 8, no. 3, pp. 551-562, 2025, doi: 10.26599/BDMA.2024.9020061.
- [15] Z. C. Liu, L. Chen, Y. J. Hu, et al., "Pe-wav2vec: A prosody-enhanced speech model for self-supervised prosody learning in tts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 1, pp. 4199-4210, 2024, doi: 10.1109/TASLP.2024.3449148.
- [16] A. Mohamed, H. Lee, L. Borgholt, et al., "Self-supervised speech representation learning: A review," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1179-1210, 2022, doi: 10.1109/JSTSP.2022.3207050.
- [17] M. Cao, F. Xu, H. Jia, et al., "A multiple interpolation algorithm to improve resampling accuracy in data triggers," *Electronics*, vol. 12, no. 6, pp. 1291, 2023, doi: 10.3390/electronics12061291.
- [18] A. S. Parihar, D. Varshney, K. Pandya, et al., "A comprehensive survey on video frame interpolation techniques," *The Visual Computer*, vol. 38, no. 1, pp. 295-319, 2022, doi: 10.1007/s00371-020-02016-y.
- [19] A. Fatwanto, F. Zamakhsyari, and R. Ndungi, "A systematic literature review of BERT-based models for natural language processing tasks," *JURNAL INFOTEL*, vol. 16, no. 4, pp. 713-728, 2024, doi: 10.20895/infotel.v16i4.1206.
- [20] A. C. Mazari, N. Boudoukhani, and A. Djeflal, "BERT-based ensemble learning for multi-aspect hate speech detection," *Cluster Computing*, vol. 27, no. 1, pp. 325-339, 2024, doi: 10.1007/s10586-022-03956-x.
- [21] J. Wu, X. Wang, X. Gao, et al., "On the effectiveness of sampled softmax loss for item recommendation," *ACM Transactions on Information Systems*, vol. 42, no. 4, pp. 1-26, 2024, doi: 10.1145/3637061.
- [22] H. Kang, T. Lv, C. Yang, et al., "Multihead-res-se residual network with attention for human activity recognition," *Electronics*, vol. 13, no. 17, pp. 3407, 2024, doi: 10.3390/electronics13173407.
- [23] C. Cao, J. Huang, M. Wu, et al., "A multivariate time series prediction method based on convolution-residual gated recurrent neural network and double-layer attention," *Electronics*, vol. 13, no. 14, pp. 2834, 2024, doi: 10.3390/electronics13142834.
- [24] X. Yuan, X. Shen, S. Mehta, et al., "Structure injected weight normalization for training deep networks," *Multimedia Systems*, vol. 28, no. 2, pp. 433-444, 2022, doi: 10.1007/s00530-021-00793-7.
- [25] L. Huang, J. Qin, Y. Zhou, et al., "Normalization techniques in training dnns: Methodology, analysis and application," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 8, pp. 10173-10196, 2023, doi: 10.1109/TPAMI.2023.3250241.
- [26] M. Kohler, J. Cho, and A. Krzyżak, "On the rate of convergence of an over-parametrized deep neural network regression estimate with ReLU activation function learned by gradient descent," *Electronic Journal of Statistics*, vol. 19, no. 2, pp. 4829-4866, 2025, doi: 10.1214/25-EJS2444.
- [27] H. Peng, B. Jiang, Z. Mao, et al., "Local enhancing transformer with temporal convolutional attention mechanism for bearings remaining useful life prediction," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, no. 1, pp. 1-12, 2023, doi: 10.1109/TIM.2023.3291787.
- [28] M. Lu and S. S. Shin, "Intelligent Temperature Control System Utilizing Gated Recurrent Units and the Internet of Things," *Journal of the Digital Content Society*, vol. 25, no. 11, pp. 3419-3429, 2024, doi: 10.9728/dcs.2024.25.11.3419.
- [29] J. Sun, Z. Wei, and X. Liu, "GRU-based model-free adaptive control for industrial processes," *Neural Computing and Applications*, vol. 35, no. 24, pp. 17701-17715, 2023, doi: 10.1007/s00521-023-08652-4.
- [30] Z. Pan, Z. Gu, X. Jiang, et al., "A modular approximation methodology for efficient fixed-point hardware implementation of the sigmoid function," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 10, pp. 10694-10703, 2022, doi: 10.1109/TIE.2022.3146573.
- [31] C. Martinelli, A. Coraddu, and A. Cammarano, "Approximating piecewise nonlinearities in dynamic systems with sigmoid functions: advantages and limitations," *Nonlinear Dynamics*, vol. 111, no. 9, pp. 8545-8569, 2023, doi: 10.1007/s11071-023-08293-1.
- [32] W. Ali, S. Vascon, T. Stadelmann, et al., "Hierarchical global attention pooling for graph classification," *Pattern Recognition Letters*, vol. 186, no. 1, pp. 71-77, 2024, doi: 10.1016/j.patrec.2024.09.009.
- [33] K. Ma, C. Tang, W. Zhang, et al., "DC-CNN: Dual-channel Convolutional Neural Networks with attention-pooling for fake news detection," *Applied Intelligence*, vol. 53, no. 7, pp. 8354-8369, 2023, doi: 10.1007/s10489-022-03910-9.
- [34] Q. Jiang, L. Zhu, C. Shu, et al., "An efficient multilayer RBF neural network and its application to regression problems," *Neural computing and Applications*, vol. 34, no. 6, pp. 4133-4150, 2022, doi: 10.1007/s00521-021-06373-0.
- [35] M. N. B. Adnan, W. M. A. W. Ahmad, N. A. Rahman, et al., "A robust hybrid methodology between applied linear regression model (alrm) and multilayer perceptron (mlp)," *Bangladesh Journal of Medical Science*, vol. 22, no. 1, pp. 38-46, 2023, doi: 10.3329/bjms.v22i1.61850.