Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6, 2536-2552
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i6.8431

The application of discourse analysis technology based on artificial intelligence in high school English teaching

Hongyu Liu[1,2], Marlina Binti Lamal[1*]
[1]School of Languages, Literacies and Translation, Universiti Sains Malaysia, Penang 11800, Malaysia; liuhongyu@student.usm.my (H.L.) jmarlina@usm.my (M.B.L.).
[2]High School Affiliated to Guangzhou University, Guangzhou 511400, Guangdong, China.

Abstract: In response to the problems existing in the application of AI (Artificial Intelligence) discourse analysis technology in high school English teaching, such as insufficient context adaptability of the automatic feedback system and weak teacher-student-AI collaboration mechanisms, this paper proposes an innovative framework of dynamic context modeling and multimodal collaborative drive. By integrating multimodal data streams of voice, text, facial expressions, and body movements in classroom scenes, a cross-modal feature dynamic fusion model is constructed to capture the semantic associations and emotional states of teacher-student interactions in real time. Based on deep reinforcement learning, a feedback algorithm with adaptive state perception capabilities is designed to periodically update the classification parameters of teaching scenes. The teacher's experience rules and AI analysis results are simultaneously embedded in the system decision-making layer through knowledge distillation technology to form a closed-loop mechanism for human-computer collaborative optimization. The experimental results show that the proposed model performs best in multiple indicators. The teacher response delay is the shortest, only 2.19 seconds; the student interaction density is the highest, reaching 14 times per minute; the student emotion scores are concentrated in a high and stable range of 3.4 to 4.3; the average score is 80.83 points. Class participation and satisfaction are also the highest, reaching 69.53% and 3.51 points, respectively, proving the advantages of the model in improving teaching effectiveness.

Keywords: *Artificial intelligence, Discourse analysis, High school English teaching, Multimodal fusion, Reinforcement learning feedback.*

1. Introduction

Currently, the application of AI in the field of education has gradually become an important direction of research and practice [1, 2]. In the field of English education, using AI to enhance teaching effectiveness and improve students' language ability has always been a hot topic for scholars and educators [3, 4]. Traditional teaching models rely on teachers' experience and students' feedback. However, with the diversification and complexity of the educational environment, a single teaching method is increasingly unable to meet the needs of different students [5, 6]. Therefore, AI-based personalized teaching has become an important way to improve teaching quality [7, 8].

In the field of discourse analysis, AI can conduct an in-depth analysis of classroom language, identify students' semantic understanding, emotional state, and language expression, and provide a real-time basis for teaching feedback, thereby optimizing the teaching process [9, 10]. Some research focuses on how to analyze students' language through technologies such as speech recognition and natural language processing [11] and assist teachers in formulating teaching strategies through feedback systems. Discourse analysis technology can interpret students' expressions from the text and speech levels [12] but in practical applications, how to effectively combine teachers' experience and contextual

History: Received: 15 April 2025; Revised: 13 June 2025; Accepted: 17 June 2025; Published: 28 June 2025
* Correspondence: jmarlina@usm.my

factors and how to establish an efficient interaction mechanism are still important issues that need to be addressed. Many existing intelligent feedback systems have limitations, such as insufficient adaptability and inaccurate context understanding, which results in their inability to maximize their role in the teaching process.

This paper implements a framework for applying AI-based discourse analysis technology to high school English teaching, aiming to address the limitations of existing teaching feedback systems. By building a dynamic context modeling and multimodal collaborative, driven framework, this paper uses cross-modal feature fusion technology to capture semantic associations and emotional changes in teacher-student interactions in real time and uses deep reinforcement learning to design an adaptive feedback mechanism to periodically update the teaching scene classification parameters, so that the system can better cope with the changing classroom environment. At the same time, the teacher's experience rules are effectively embedded in the decision-making layer through knowledge distillation technology, further enhancing the decision-making ability of AI and finally realizing a closed-loop mechanism of human-machine collaborative optimization. This study not only provides a new AI support solution for high school English teaching but also provides strong theoretical support and a practical basis for the development of future educational technology.

2. Related Work

With the development of AI technology, its application has expanded to personalized learning [13, 14] and automated feedback systems [15, 16]. These systems can provide personalized suggestions and real-time feedback based on students' learning situations, but most of the current AI teaching systems still have problems, such as poor adaptability and inflexible feedback. Especially in a highly dynamic classroom environment, how to deal with complex contexts and emotional changes is still a challenge.

In terms of discourse analysis, the application of artificial intelligence technology has brought new opportunities for teaching [17, 18]. Traditional discourse analysis mainly focuses on the grammatical and semantic structure of the text, while modern AI discourse analysis pays more attention to the emotional and contextual information in the language [19, 20]. Through natural language processing and sentiment analysis, AI can deeply understand students' language behavior, emotional state, and level of understanding. However, existing discourse analysis technology still faces problems of accuracy and real-time performance, especially in the classroom, where language is not just a simple transmission of text or voice but also includes students' emotions and body language. Therefore, comprehensively analyzing students' expressions and adapting to contextual changes in real-time interaction is difficult in current discourse analysis research.

Multimodal data fusion has been an important direction in the field of AI in recent years [21, 22]. In educational applications, multimodal fusion technology can combine multiple information sources such as speech, text, expression, and body language to fully understand students' learning status [23, 24]. For example, AI systems can obtain students' emotional state and learning engagement by analyzing their speech and text content. However, the current application of multimodal data fusion technology in education still faces problems, such as difficulty in data integration and weak feedback capabilities. In the classroom interactive environment, how to efficiently integrate multimodal information and adjust teaching strategies in real-time remains a challenge [25, 26].

In terms of human-computer collaborative teaching, recent studies have shown that AI can be used as an auxiliary tool for teachers to help improve teaching effectiveness [27, 28]. AI can not only analyze students' performance in real-time and provide personalized feedback but also assist teachers in making teaching decisions. Existing research focuses on how to use AI to support personalized teaching [29-30] and automated feedback [29] but these systems usually rely on fixed rules and models, lack sufficient flexibility and adaptability, and cannot effectively deal with complex interactions in the classroom. How to achieve collaborative cooperation between teachers, students, and AI in a changing teaching environment and improve the adaptability and response speed of AI systems is still a difficult point in current research.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

In response to these problems, this paper proposes an artificial intelligence-based discourse analysis framework to address the shortcomings of existing AI teaching systems. By integrating multimodal data such as speech, text, expressions, and body language, this paper constructs a cross-modal feature dynamic fusion model that can capture students' semantic and emotional states in real-time. The adaptive feedback mechanism designed by deep reinforcement learning enables the AI system to dynamically adjust according to changes in the teaching scenario and embeds teacher experience into the decision-making layer through knowledge distillation technology to optimize teaching decisions. These innovations effectively improve the contextual adaptability and interactivity of the system, providing a new intelligent tutoring solution for high school English teaching.

3. Construction of an Interactive English Teaching System Driven by Intelligent Discourse Modeling

3.1. Multimodal Acquisition Strategies for Interactive Features of Discourse in High School English Classes

In this study, multimodal data collection in high school English classes adopt a variety of efficient equipment and technologies to ensure the accurate acquisition of voice, text, facial expression, and body movement data, facilitating subsequent interaction analysis and emotion recognition.

The collection of text data uses speech recognition technology to convert the voice data collected in the classroom into text in real-time. During each speech, the system stores and updates the converted text synchronously, and every language input of teachers and students in the teaching interaction is accurately recorded. After the text is converted, the text processing module is used to timestamp and analyze the transcribed text to ensure that the emotional tendency and semantic information of each speech can be reflected in a timely and accurate manner.

The facial expression data is collected through high-definition cameras set up in the classroom, which monitor the facial expressions of students and teachers in real-time. The facial expression recognition system uses key point detection algorithms to extract key feature points of the face and analyze the changes in facial expressions based on this.

The collection of body movements uses a combination of depth sensors and infrared cameras. Multiple sensors deployed in the classroom can capture all body movements. The system analyzes students' participation status and behavioral patterns in classroom interactions by capturing the body postures, gestures, and movement trajectories of students and teachers. The sensor tracks body movements in real-time and converts them into three-dimensional coordinate points. The motion capture algorithm reconstructs the trajectory of these coordinate points and identifies students' standing, sitting, raising hands, walking, and other behavioral actions. After being annotated, the body movement data is stored synchronously with the voice and expression data to ensure that they have a unified time label in subsequent analysis. The multimodal data format is shown in Table 1.

Table 1.
Multimodal data formats.

Modality Type	Collection Method	Data Format	Sampling Frequency
Voice	Microphone	.wav	16kHz
Text	Transcription + PPT	.txt/.docx	Real-time Transcription
Facial Expression	Camera	.json	30 fps
Motion	Sensor and Camera	.json	30 fps

All data is transmitted to the central processing system via real-time data streams to ensure the synchronization and integration of multimodal data. The timestamps of speech, text, facial expressions, and body movements are precisely aligned to ensure that the relationship between various types of data can be analyzed within the same time window.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

3.2. Cross-modal Semantic Fusion and Dynamic Context Modeling Mechanism

In this study, the implementation process of cross-modal semantic fusion and dynamic context modeling mechanism is based on a deep learning model, using the Transformer architecture to fuse and process multimodal data. The process specifically includes four core steps: data preprocessing, modal feature extraction, cross-modal fusion, and dynamic context modeling.

The speech signal is extracted through a short-time Fourier transform and MFCC (Mel Frequency Cepstrum Coefficient) to generate a feature matrix $X^{(v)} \in \mathbb{R}^{T\times F}$, where T represents the number of time frames, and F is the frequency dimension. This feature is extracted through CNN (Convolutional Neural Networks) to extract the time-frequency local features:

$$H^{(v)} = \mathrm{CNN}\left(X^{(v)}\right), \quad H^{(v)} \in \mathbb{R}^{T\times d} \ (1)$$

d is the feature map dimension. Then, LSTM (Long Short Term Memory) captures the temporal dependency and obtains the hidden state sequence:

$$S^{(v)} = \mathrm{LSTM}\left(H^{(v)}\right), \quad S^{(v)} \in \mathbb{R}^{T\times d} \ (2)$$

The text data input is a transcribed sentence sequence $W = \{w_1, w_2, \ldots, w_N\}$, which is encoded through the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to obtain semantic embedding:

$$E^{(l)} = \mathrm{BERT}(W), \quad E^{(l)} \in \mathbb{R}^{N\times d} \ (3)$$

The expression data consists of a sequence of facial key point images, which are input into the ResNet deep convolutional network to extract spatial features:

$$F^{(f)} = \mathrm{ResNet}\left(I^{(f)}\right), \quad F^{(f)} \in \mathbb{R}^{T\times d} \ (4)$$

The time series of key points of limb movements $J = \{j_1, j_2, \ldots, j_T\}$ is input into the time series graph convolutional network to extract dynamic behavior features:

$$G^{(g)} = \mathrm{STGCN}\left(I^{(f)}\right), \quad G^{(g)} \in \mathbb{R}^{T\times d} \ (5)$$

After feature extraction, each modal data is aligned through precise time alignment technology to ensure that the timestamps of different modalities are consistent and avoid information loss caused by timing misalignment. After time alignment, the data streams of speech, text, expression, and body movements enter the cross-modal fusion stage. Each modal feature is linearly mapped and embedded into the same dimensional space, then standardized through position encoding and input into the Transformer model. Transformer's self-attention mechanism calculates the correlation between each modality through multi-head self-attention, evaluates the relative importance of each modality at the current moment, and fuses the features of each modality through weighted summation. The self-attention mechanism can capture the global correlation between modalities at each layer, effectively avoiding the problem of modality separation in traditional methods.

For the features obtained after cross-modal fusion, dynamic context modeling further enhances the model's ability to perceive context. The context of classroom interaction is dynamically changing, and the emotions, tone, intonation, and behavior of teachers and students change over time. Therefore, the model needs to dynamically adjust the semantic and sentiment analysis results at each moment according to the changes in the time series. Based on the time series modeling capability of Transformer, the model can capture and model the time dependencies in classroom interactions through the contextual information of historical data. The output at each moment depends not only on the input at the current moment but also on the state at the previous moment. In this way, the model can adjust the weight of different modal data at the current moment according to the emotional state and semantic information at the previous moment so that the model gives appropriate attention to different modalities in a dynamic context.

The application of emotion perception strengthens the model's response to emotional fluctuations in the classroom. During classroom interaction, the emotional and semantic expressions of teachers and students fluctuate. The Transformer's self-attention mechanism can capture these fluctuations and model them. The model can adjust the current moment's emotion analysis strategy based on historical

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

emotion information in a timely manner, thereby more accurately identifying emotional changes in the classroom, which in turn affects subsequent teaching decisions and feedback.

3.3. Design of Adaptive Feedback Mechanism

The adaptive feedback mechanism is built based on the Proximal Policy Optimization (PPO) algorithm, and its core goal is to select the optimal feedback strategy for real-time context and modal state in a dynamic classroom environment. PPO achieves stability and efficiency in the policy update process through the policy gradient optimization mechanism, which is suitable for real-time policy selection in high-dimensional state space. The mechanism uses temporal situational states and multi-dimensional emotion vectors to form the state space, and the action space is defined as the set of feedback behaviors that the system can choose, including specific feedback methods such as language guidance, expression imitation, rhythm adjustment, and task reconstruction.

The state representation vector is composed of the semantic context features of the current moment, the change in intonation frequency, the facial expression embedding vector, and the historical interaction behavior. The emotional state is converted into continuous variables (such as pleasure, tension, and engagement) through real-time analysis by an external model and normalized and incorporated into the state vector. In order to improve the temporal sensitivity of state recognition, a gated recurrent unit is introduced to compress and model the state sequence so that PPO can extract long-term dependencies from the past several steps of the state.

The action selection is output by the policy network. The policy network structure is a two-layer feedforward neural network with 128 and 64 nodes in each layer, and the activation function uses ReLU. The network input is the state vector, and the output is the action probability distribution. The entropy regularization term is used in the sampling process to adjust the diversity of the policy distribution and improve the system's ability to explore different feedback actions in the initial stage. The value function network structure is symmetrical and is used to estimate the expected return of each state-action pair. It also participates in the advantage function calculation as a baseline in the policy update stage.

The policy update uses a clipping objective function to limit the ratio of the new and old policies within a certain range to prevent performance degradation caused by large policy updates. In each epoch, the policy network and the value network are updated for multiple rounds. The optimization goal is to maximize the expected reward, minimize the value function error, and minimize the negative value of the policy distribution entropy to control the policy stability.

The reward function design adopts the modality perception mechanism and dynamically assigns values based on the teacher-student behavioral responses after system feedback. Short-term positive rewards come from indicators such as the migration of students' facial expressions to positive emotions, the stabilization of speech speed and tone, and the reduction of interactive response delays; negative rewards are triggered by negative emotional migration of students, lack of response after feedback, or decreased participation. In order to solve the sparse reward problem, an intermediate reward item based on the gradient of emotional fluctuations is introduced, and the direction of strategy adjustment is guided by calculating the rate of change of emotional variables at consecutive moments.

Feedback behavior decision-making uses a single-step action as the basic unit, but considering that classroom behavior has the characteristics of sequential decision-making, the policy network uses the time difference method to estimate the expected return during the training phase. After the system completes a teaching unit, it uses the trajectory playback mechanism to update the policy buffer and evaluate the policy performance changes in the form of a sliding window.

3.4. Knowledge Distillation Strategy Embedded with Teachers' Experience Rules

The integration of teacher experience rules and AI model reasoning results is based on a guided knowledge distillation framework, which consists of three parts: teacher rule model, student strategy model, and distillation guidance mechanism. Its architecture is shown in Figure 1.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

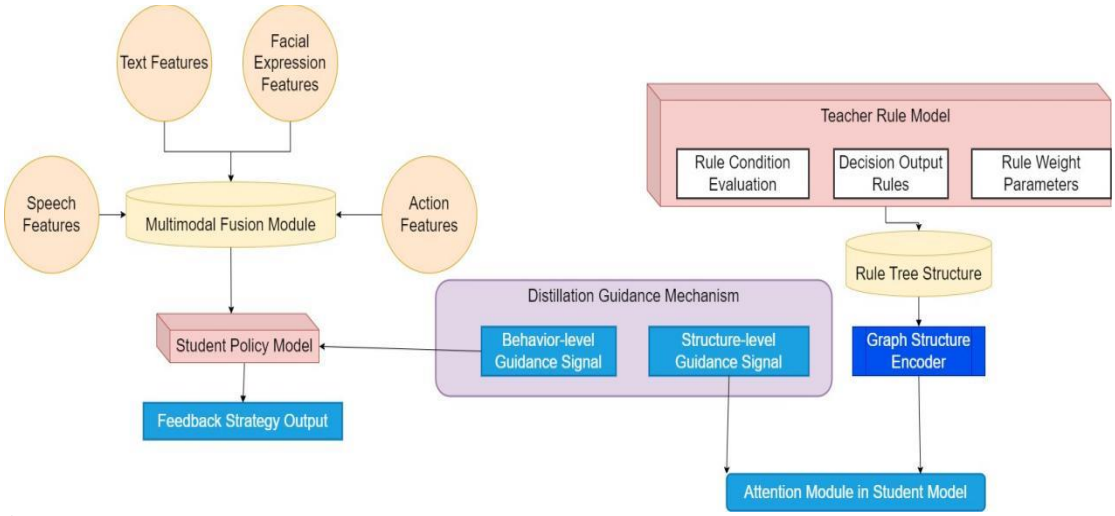

Figure 1.
The overall architecture of the knowledge distillation strategy guided by teacher rules.

The teacher rule model is centered on a structured decision template, which encodes the feedback experience accumulated by human teachers in the classroom interaction process into a set of logical rules. The rules cover the classification of discourse comprehension errors, the matching relationship between the student's emotional state and feedback intensity, and the weight of the interaction stage on the feedback strategy. Each rule consists of three parts: conditional judgment, decision output, and weight parameter. It is organized in a tree structure so that the rules can be converted into a graph structure for neural network embedding and processing.

The rule model does not directly participate in forward reasoning but is used as a teacher model to guide the student model to learn feedback generation strategies. The student model is built based on the Transformer structure, accepts cross-modal fusion of speech, text, expression, and action representation as input, and outputs feedback behavior suggestions and their confidence after multi-layer encoding and state estimation modules. During the training phase, the distillation module uses the behavior labels and weight information provided by the teacher rule model as soft targets to jointly optimize the strategy distribution and attention allocation mechanism of the student model so that it gradually approaches the feedback preferences and structural characteristics of the teacher rules.

The rule distillation process includes two types of guidance signals: behavioral level and structural level. At the behavioral level, the teacher model generates feedback suggestions and probability distribution for each training sample scenario based on the rule set. The student model needs to fit the distribution at the same time during training and gradually approach the behavioral preferences reflected by the teacher's experience while maintaining the autonomous exploration space. At the structural level, the conditional path of each rule in the teacher model is encoded as a vector and trained to align with the output of the Transformer attention layer in the student model. This strengthens the student model's explicit attention to rule logic when generating feedback and improves the consistency of its decision-making path with the teacher's experience.

Rule embedding is implemented using a graph structure encoder, which converts the node representation in the teacher's rule graph into a vector and inputs it into the attention module of the student model to form a rule-guided item, which is embedded in the attention allocation stage of the attention calculation formula to guide the student model to consider the key factors shown by the teacher model when making decisions among multimodal features. Under this guidance, the student model not only learns specific feedback behavior samples but also absorbs the strategic basis of how the teacher model makes judgments and trade-offs in different teaching contexts, thereby having stronger generalization and explanation capabilities.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

3.5. Closed-Loop Optimization Architecture for Intelligent Human-Machine Collaboration

The overall architecture of the intelligent human-machine collaborative system consists of a perception module, a decision module, an execution module, and an evaluation module, forming a closed-loop structure covering the entire process of collection, analysis, feedback, and regulation. In the system operation process, the perception module receives multimodal input in real-time and encodes it into a context-state representation. The decision module generates feedback actions based on the current state. The execution module completes the specific feedback push. The evaluation module quantitatively analyzes the feedback effect and uses the evaluation results as environmental rewards to update the strategy network, forming a dynamic feedback path driven by reinforcement learning. The closed-loop optimization architecture of the intelligent human-machine collaborative feedback system is shown in Figure 2.

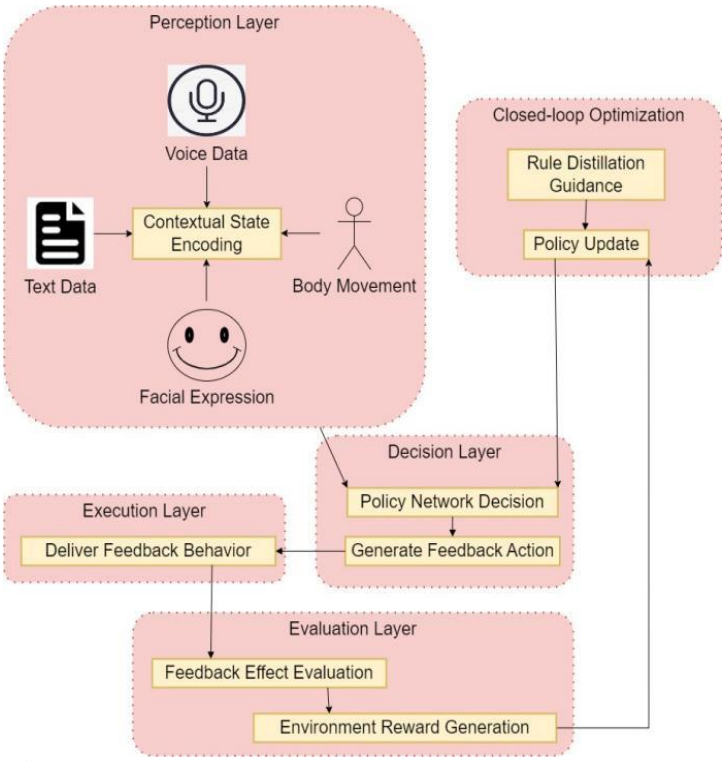

Figure 2.
Closed-loop optimization architecture of intelligent human-machine collaborative feedback system.

The entire system is executed around the teaching interaction cycle, continuously receiving student status changes and teaching feedback performance. After each interaction cycle, the cycle update scheduling mechanism calls the local distillation training and strategy re-optimization process, dynamically adjusts the strategy parameters and rule embedding representation, and enhances the system's adaptability to variant contexts and strategy drift. The system's built-in strategy degradation discrimination mechanism can automatically trigger high-frequency fine-tuning or low-frequency distillation training according to changes in feedback stability during the teaching process, improving the consistency of feedback behavior and the quality of teaching response. The integrated parameter update channel linkage rule embedding optimization and strategy generation network training constitutes a closed-loop optimization mechanism with periodic retraining, adaptive fine-tuning, and experience incremental distillation as the core, realizing the long-term stable evolution and short-term

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

situational adaptation of the feedback strategy, and supporting the continuous and effective operation of the AI decision-making system in high-frequency and high-complexity classroom interaction scenarios.

4. Experimental Design and Implementation

4.1. Multimodal High School English Classroom Corpus Collection

The experimental data comes from a one-semester English course in two public high schools, covering 14 teaching units and a total of 42 natural teaching scenes, involving 6 English teachers and 93 students. The classroom is configured as a standard multimedia classroom without special modifications. The audio and video data is provided by a Blue Yeti microphone (USB interface, sampling rate 48kHz), a Logitech Brio 4K camera (30fps), a Xiaomi DeepSense camera, and OBS Studio screen recording software.

The speech signal is processed by pydub and librosa libraries, and short-time Fourier transform and 13-dimensional MFCC feature extraction are performed. The first 60 frames are retained to form the audio feature matrix. The video image is frame-cut, and the facial feature extraction is completed by the media pipe.face_mesh module, which outputs the coordinates of 478 facial mesh key points; the text is transcribed in real-time by iFlytek speech recognition SDK (Software Development Kit), and the transcription results are manually reviewed twice, with a sentence-level error correction accuracy of more than 98%; the body movements are extracted by the mediapipe .pose a model to extract 33 skeleton key point coordinates, and the frame rate is limited to 15fps to match the device processing capability.

Multimodal data is timestamped by a custom Python script. All modalities are based on speech and dynamic time warping is used to adjust frame synchronization and output structured interaction sequences. Data annotation is completed using the labelstudio platform. Emotional labels are divided into four categories: high concentration, high pleasure, neutral, and negative, based on clues such as teacher tone, expression, and student reaction. Pragmatic behavior annotation includes seven labels: teacher questions, explanations, feedback, student responses, silence, and active questions. The annotation team consists of four people, including teaching researchers and experienced teachers. A double-label consistency mechanism is used, and the Kappa consistency is 0.78.

All features are normalized by Z-score, and data enhancement methods include sentence length perturbation (±15%), frame interpolation, and volume perturbation (±3dB). Finally, a multimodal sample structure is generated for training: speech (60×13), text (sentence-level BERT embedding, 768 dimensions), expression (64-dimensional compressed vector), and action (33×2 skeleton coordinates).

4.2. Model Training and Parameter Setting

The cross-modal fusion model training is implemented using the PyTorch framework and is completed in a common commercial hardware environment, including an NVIDIA RTX 3060 GPU (12GB video memory), an Intel i7-12700F processor, and 32GB RAM. The model input is a multimodal feature sequence, which comes from speech (CNN+LSTM output), text (BERT embedding), expression (ResNet extraction), and body movement (ST-GCN encoding). All modal features are linearly mapped to 128 dimensions and then input into the Transformer module after position encoding. The Transformer adopts a two-layer structure; each layer contains 4 attention heads, and the hidden dimension is 256. The Dropout layer is applied to the model to suppress overfitting, and the ratio is set to 0.3.

The context modeling part is embedded in a single-layer GRU (Gated Recurrent Unit) with 128 hidden units, which is used to learn the temporal dependencies in the context of continuous teaching. The training optimizer is Adam, the initial learning rate is set to 1e-4, the number of training rounds is 40, and the multi-label cross-entropy loss function is used for error backpropagation. The early stopping mechanism dynamically determines the convergence based on the change of the F1-score of the validation set.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

The PPO feedback strategy adopts a dual network structure, with the policy network and the value function network configured symmetrically. The policy network input is the current fused state vector and the last three rounds of feedback embedding, totaling about 256 dimensions, with a two-layer MLP (Multilayer Perceptron) structure, with 128 and 64 hidden units, respectively, the activation function is ReLU, and the output layer is connected to the softmax distribution. In PPO training, the shear coefficient is set to 0.2, the GAE (Generalized Advantage Estimation) parameter is set to 0.95, the discount factor is set to 0.9, and the entropy regularization term coefficient is set to 0.01. The number of interaction steps per round is set to 50, the total number of episodes is set to 10, the policy buffer capacity is set to 1000, and the time difference method is used to estimate the return. The strategy update saves the weight every 2 rounds, and the convergence criterion is that the average reward increases by less than 1% for three consecutive rounds. The key parameter settings of the model are shown in Table 2.

Table 2.
Key parameter settings of the model.

Module	Parameter	Setting Value
Transformer Encoder	Number of Layers	2
Attention Heads	Per Layer	4
Modal Mapping Dimension	Unified Dimension for All Modalities	128
GRU Hidden Units	Dynamic Context Modeling	128
Initial Learning Rate	Adam Optimizer	1e-4
Dropout Rate	Applied Throughout Model	0.3
Policy Network Structure	MLP Hidden Layers	[128, 64]
PPO Clipping Coefficient	ε	0.2
Discount Factor	γ	0.9
GAE Parameter	λ	0.95

4.3. Teaching Experiment

93 students are randomly divided into three groups, corresponding to the traditional teaching group (Group 1), the general discourse analysis assisted teaching group (Group 2), and the multimodal intelligent discourse analysis assisted teaching group based on this paper (Group 3). All experiments are carried out in a standard multimedia classroom environment, and the teaching equipment and acquisition configuration remained consistent. The traditional teaching group adopts conventional high school English teaching content and methods without any artificial intelligence auxiliary tools. The general discourse analysis assisted teaching group provides assisted teaching based on unimodal text discourse analysis technology. The system only uses transcribed text data to analyze and provide feedback on classroom discourse. The feedback strategy is relatively static and lacks cross-modal context perception capabilities. This group fully applies the multimodal discourse analysis model and adaptive feedback mechanism based on voice, text, facial expressions, and body movements to capture the semantic and emotional changes of teachers and students in real-time and realize intelligent teaching feedback driven by dynamic context. The three groups of experiments last for one semester, covering the same 14 teaching units to ensure the consistency of teaching content and teaching progress. Each group of experiments is equipped with English teachers with the same qualifications to teach. The teachers maintain their original teaching style in the traditional group and provide feedback according to the system suggestions in the auxiliary teaching group.

In addition to the formal model training and teaching experiments, two ablation experiments are designed to verify the impact of the key modules of the system on semantic recognition and teaching feedback quality. The first ablation experiment removes the multimodal input and retains only the text corpus input. The output is used to compare the semantic recognition accuracy and analyze the recognition gain of the multimodal discourse analysis model after the introduction of modal features such as expressions and actions; the second ablation experiment targets the teacher experience distillation module. While keeping the rest of the model structure and parameters consistent, the

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

teacher rule embedding branch is removed to evaluate the performance differences of the system in recommending feedback content and adjustment strategies, respectively, to analyze the contribution of the distillation strategy to the quality of teaching decisions.

5. Conclusion

5.1. Classroom Interactivity

In order to compare and analyze the characteristics of different teaching methods in the dimension of classroom interaction, teacher response delay and student interaction density are selected as representative indicators to measure the timeliness of teaching response and the activeness of student classroom participation, respectively. The relevant distribution and confidence region are shown in Figure 3.

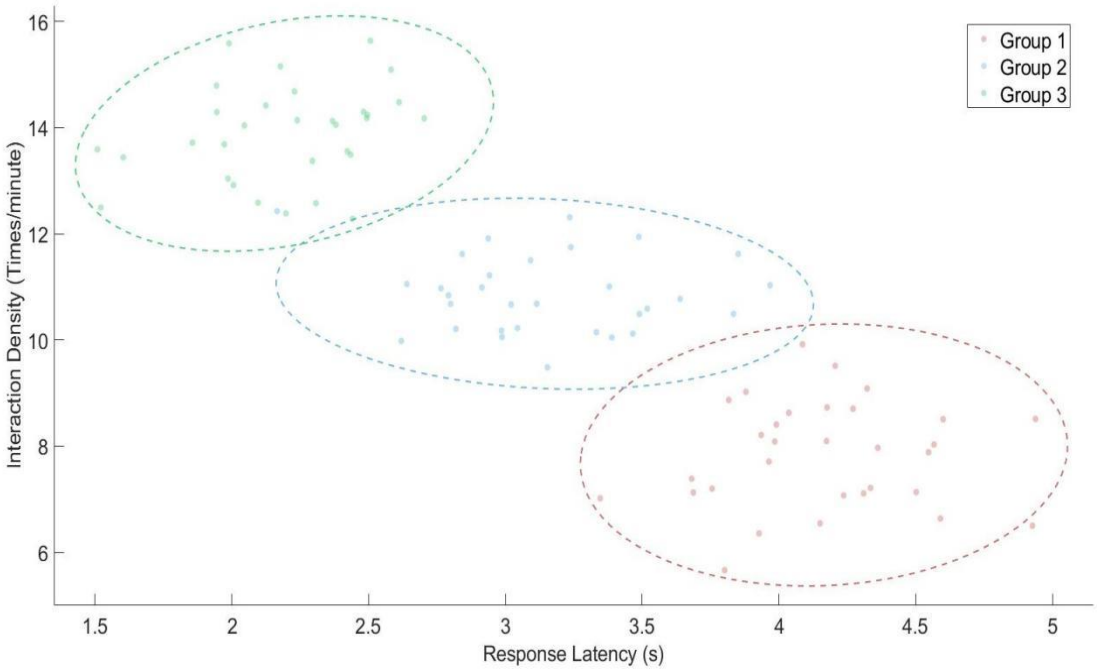

Figure 3.
Distribution of teacher response delay and student interaction density under three teaching methods

Figure 3 shows the performance differences between the three teaching methods in terms of classroom interaction. The horizontal axis is the teacher's response delay, and the smaller the value, the faster the response; the vertical axis is the student interaction density, and the larger the value, the more active the student participation. Group 1 has an average response delay of about 4.16 seconds and an average interaction density of about 8 times/minute; Group 2 has a response delay of 3.14 seconds and an interaction density of about 11 times/minute; Group 3 performs best, with a teacher response delay of 2.19 seconds and a student interaction density of 14 times/minute, and a more concentrated distribution, indicating that the teaching behavior is more efficient and the student participation is more active.

From the perspective of model mechanism, the traditional teaching model has a relatively simple form of interaction, which leads to delayed teacher response and lack of sufficient motivation for students to participate; discourse teaching applies real context and is task-driven, which improves the coherence of teacher dialogue and the frequency of interaction so that teachers respond faster and students are more active. The multimodal AI teaching method integrates intelligent mechanisms such

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

as speech recognition, emotion detection, and real-time feedback, which can dynamically capture student behavior and emotional changes and quickly make personalized responses, thereby greatly shortening the teacher's response delay and promoting high-frequency interaction.

5.2. Changes in Students'Emotional State in English Learning

In order to further explore the impact of different teaching strategies on the evolution of students' emotional states in the process of English learning, the PANAS (Positive and Negative Affect Schedule) scale, widely used in the field of emotional psychology, is used to quantitatively model students' positive emotional experiences. This experiment uses 14 teaching units as the time series dimension to collect and simulate student emotional rating data under three methods: traditional teaching, discourse-based teaching, and multimodal AI teaching. The results are shown in Figure 4.

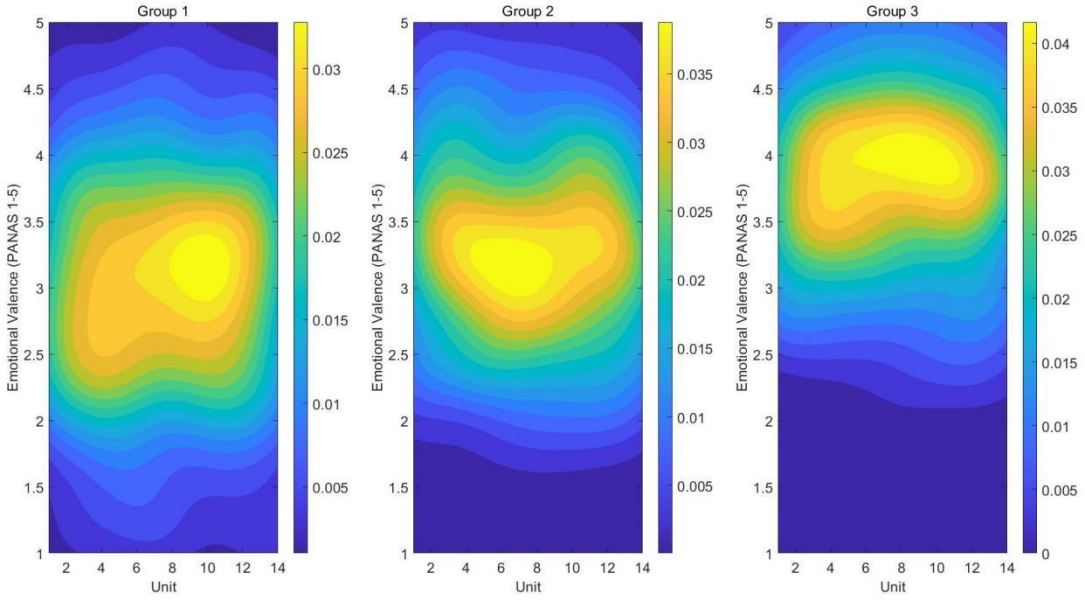

Figure 4.
Distribution of students' emotional states under different teaching modes.

As can be observed from Figure 4, Group 1 presents a relatively low-density concentrated area in terms of emotion scores, mainly distributed between 2.3 and 3.5, and the distribution is relatively discrete; Group 2 shows a trend of density moving toward high segments, indicating that students maintained a high level of emotional experience in most teaching units; Group 3 forms an obvious high-density concentrated zone in the range of 3.4 to 4.3, especially in units 5 to 12, indicating that the overall emotions of this group of students are maintained at a higher level during the learning process, and the fluctuations are relatively small.

The above differences in emotional states reflect the potential impact of different teaching methods on the stimulation of students' positive emotions. Under traditional teaching methods, emotional stimulation relies on the direct stimulation of teacher behavior and textbook content, which is easily disturbed by classroom rhythm and individual differences, thus leading to large fluctuations in emotional levels. The discourse-driven approach guides students to participate in language comprehension and generation through situational semantics, which partially improves learning motivation and emotional involvement, as reflected in the positive shift of scoring density. On this basis, multimodal AI further integrates multimodal inputs such as voice, vision, and interactive feedback to form a more accurate adaptation and regulation mechanism for students' emotional states, significantly

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

improving the continuity and stability of students' emotional experience in each teaching unit. This maintenance of positive emotions is not only conducive to stimulating interest in learning but may also reversely promote the improvement of language processing efficiency.

5.3. Comparison of Teaching Results

In order to explore the impact of multimodal teaching on students' comprehensive learning performance, the experiment compares the three dimensions of learning outcomes, classroom participation, and satisfaction, as shown in Figures 5 and 6.

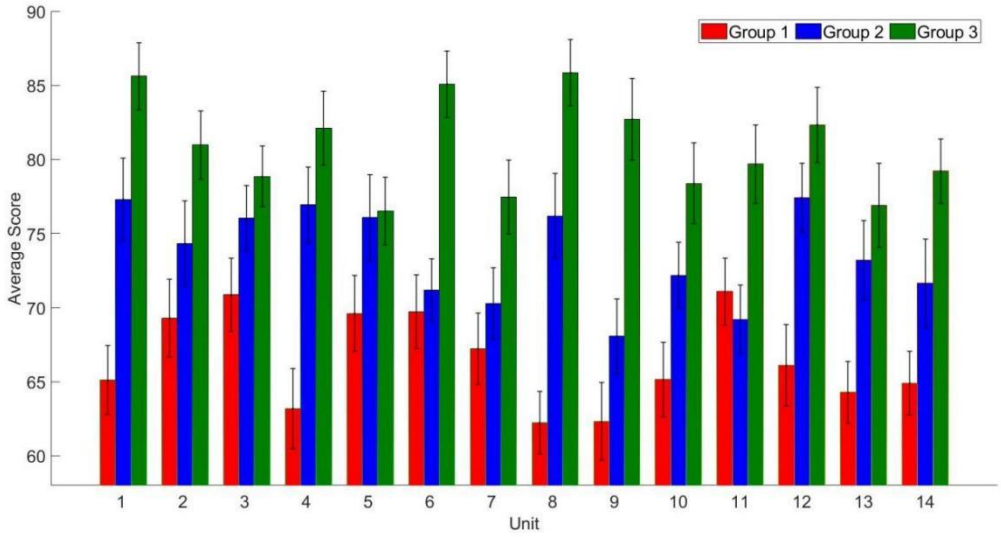

Figure 5.
Comparison of teaching results of different teaching groups.

The average scores of Group 3 in each unit are around 80.83 points. The scores of Group 2 are mainly distributed between 68.07 and 77.43 points, with an overall performance of above average, while the scores of the traditional group are mostly concentrated between 62.25 and 71.08 points.

This phenomenon of grade differentiation shows that different teaching models have different effectiveness in knowledge construction. The multimodal group uses AI to help integrate text, images, and audio, making the learning materials more perceptually advantageous and inspiring students to maintain continuous attention in each unit, thereby promoting steady improvement in grades. The task-driven approach of the discourse group helps with the logical understanding of knowledge but lacks multimodal feedback support, and its effectiveness is inferior to the former. Traditional teaching, however, makes it difficult to form a strong unit learning rhythm due to the single interactive method and delayed feedback, resulting in large fluctuations in some units. These trends clearly reflect the direct traction of teaching design on academic output.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

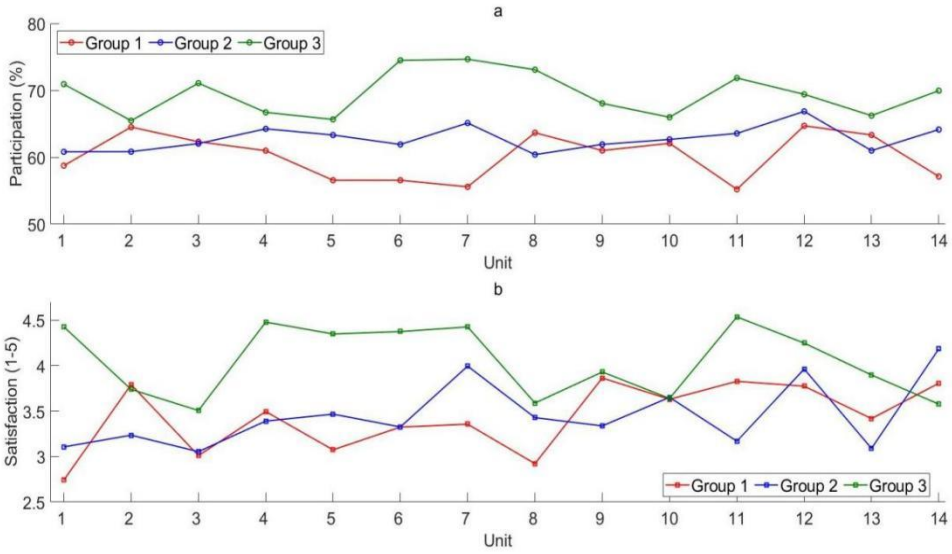

Figure 6.
Comparison of class participation and learning satisfaction.

In terms of participation, Group 3 has an overall average of 69.53%, with the highest reaching a peak of nearly 74.66%, indicating that its teaching environment can effectively mobilize students to continue to participate. Group 2 is between 60.39–66.85%; Group 1 has an average of about 60.17%. In terms of satisfaction, Group 3 is above 3.51 points, with the highest reaching 4.54 points; Group 2 fluctuates between 3.05–4.18, while Group 1 is concentrated between 2.74–3.86 points and is generally in a weaker perception range.

This set of data reveals the dual impact mechanism of teaching methods on students' subjective experience and classroom engagement. The multimodal group maintains its lead in both participation and satisfaction, indicating that its system not only enhances cognitive appeal but also provides positive feedback at the emotional level, prompting students to establish a stable learning rhythm and high satisfaction; the discourse group helps maintain a certain degree of participation in terms of structural clarity, but due to the lack of sensory stimulation and personalized feedback, its satisfaction is average; the traditional group is at a disadvantage in terms of participation willingness and satisfaction due to the lack of interaction and strategic support. These trends reinforce the potential advantages of multimodal AI teaching in stimulating students' initiative and positive emotions.

5.4. Effect of Multimodal Discourse Analysis Model on Improving Accuracy

In order to systematically evaluate the impact of multimodal information fusion on the classification performance of the model, an accuracy comparison experiment based on different input combinations is designed. The experimental setting includes four input modes: "text", "text + facial expression", "text + facial expression + action," and "full modality". The results of the first ablation experiment are shown in Figure 7.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

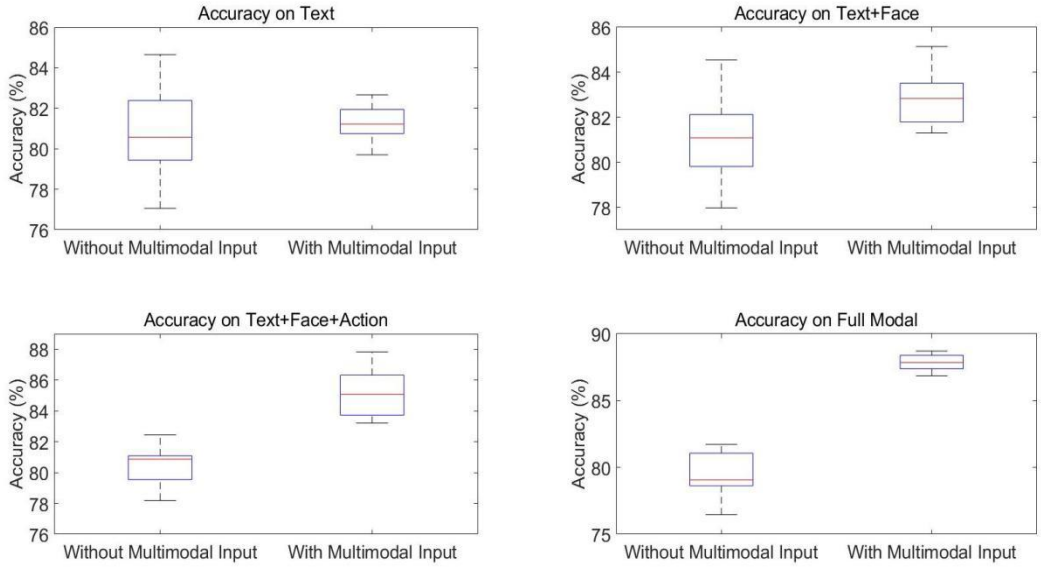

Figure 7.
Comparison of model analysis accuracy.

In all input combinations, the accuracy of the multimodal fusion model is higher than that of the non-fusion model, and as the input modality increases, the accuracy of the fusion model gradually increases. Under full-modal input, the average accuracy of the fusion model is close to 88%, while the non-fusion model remains at around 80%, showing that multimodal fusion significantly improves classification performance. The reason behind this performance improvement is mainly due to the fact that multimodal fusion can integrate different types of information, make up for the limitations of a single modality, and improve the model's expression and discrimination capabilities. Text information provides the basis for semantic content, while facial expressions and movements as visual signals capture subtle changes in emotions and behaviors, which helps enhance contextual understanding and emotion recognition, thereby improving overall recognition accuracy. The multimodal fusion mechanism effectively integrates this heterogeneous information, achieves information complementarity, reduces the impact of noise, and enables the model to capture target features more comprehensively, thereby showing better recognition results under different input combination conditions, reflecting the robustness and advantages of the fusion strategy in complex environments.

5.5. The Impact of Teacher Experience Embedding on the Quality of System Teaching Decisions

In order to evaluate the impact of teacher knowledge distillation technology on the performance of the recommendation system and decision-making module, the experiment covers two dimensions: recommendation quality and decision performance. The recommendation accuracy, relevance score, diversity index, decision consistency score, strategy adjustment response time, and decision stability index are compared to analyze the contribution and optimization effect of knowledge distillation on the overall performance of the system. The relevant statistical significance is tested by the P value to ensure the scientificity and reliability of the results. Table 3 shows the results of the experiment.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

Table 3.
Comparison of teaching decision quality.

Metric Category	Metric Name	With Teacher Knowledge Distillation	Without Teacher Knowledge, Distillation	P-value
Recommendation Quality	Recommendation Accuracy (%)	87.3	79.8	<0.05
	Recommendation Relevance Score	4.25 / 5	3.68 / 5	<0.05
	Recommendation Diversity Index	0.68	0.52	<0.05
Decision Performance	Decision Consistency Score	0.91	0.75	<0.05
	Strategy Adjustment Response Time (minute)	1.85	3.42	<0.05
	Decision Stability Index	0.89	0.72	<0.05

After enabling teacher knowledge distillation, the recommendation accuracy rate is significantly improved to 87.3%, the recommendation relevance score is also improved from 3.68 to 4.25, and the diversity index increases from 0.52 to 0.68, indicating that the recommendation results are not only more accurate but also more diverse. In terms of decision performance, the decision consistency score increases from 0.75 to 0.91, the strategy adjustment response time is shortened from 3.42 minutes to 1.85 minutes, and the decision stability index increases from 0.72 to 0.89. All indicators show significant statistical differences (P values are all less than 0.05), reflecting the comprehensive improvement of system performance by knowledge distillation.

The reason behind these data is mainly due to the fact that teacher knowledge distillation guides the student model to learn the rich knowledge of the teacher model, realizes the effective transfer and fusion of knowledge, and thus enhances the generalization ability and prediction accuracy of the model. The distillation process not only optimizes the relevance and diversity of the recommendation results but also improves the response speed and stability of the decision-making module, reflecting the improvement of the model's adaptability and execution efficiency in a dynamic environment. This mechanism promotes the coordinated optimization of the recommendation and decision-making systems by strengthening the capture of potential patterns and complex features and further promotes the improvement of the overall performance of the system.

5.6. Delayed Post-test Analysis of Students' After-class English Usage Behavior and Learning Outcomes

In order to explore the impact of different teaching modes on students' after-class English use behavior and learning outcomes, each group of students took a one-month delayed post-test after completing classroom learning to re-evaluate their mastery of the same knowledge points. At the same time, the students' after-class English use frequency, active output ratio, and actual practice time are continuously tracked. Through the collection and comparison of multi-dimensional indicators, the impact of different teaching interventions on students' language application habits and learning continuity is revealed. The results are shown in Table 4.

Table 4.
Comparison of delayed post-test scores and post-class English usage behaviors.

Group Description	Group 1	Group 2	Group 3
Sample Size	31	31	31
Delayed Post-Test Score	71.4	76.9	83.7
Weekly Post-class English Usage Frequency	2.1	3.4	5.2
Proportion of Active English Output (%)	24.3	35.6	51.8
Average Duration of English Practice (minutes/week)	30	45	70
Frequency of Using English Apps (times/week)	1.8	2.9	4.7
Number of English Conversations per week	1.2	2.1	3.8

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431

From the delayed post-test results, group 3 performs best, with an average score of 83.7 points, higher than group 1's 71.4 points and group 2's 76.9 points. In terms of the frequency of English use after class, group 3 uses English an average of 5.2 times per week, much higher than group 1's 2.1 times and group 2's 3.4 times; the proportion of active English output also increases significantly, reaching 51.8% for group 3, while group 1 and group 2 are 24.3% and 35.6%, respectively. In addition, group 3 students lead in weekly English practice time, frequency of using English learning applications, and number of English conversations, reaching 70 minutes, 4.7 times, and 3.8 times, respectively, showing more active and systematic after-class language application behavior.

These data reflect that multimodal intelligent discourse analysis-assisted teaching has a significant effect on promoting students' language learning continuity and active participation. The multimodal fusion intelligent feedback mechanism can accurately capture students' context and emotional state and can adjust teaching strategies in an individualized manner, enhance the immersion and pertinence of the learning experience, and thus stimulate students' willingness to actively practice and apply English in practice. Compared with the traditional and single discourse analysis-assisted mode, the closed-loop feedback and optimization mechanism formed by the intelligent model inside and outside the classroom effectively improves students' language output ability and communication confidence, promotes the internalization of learning motivation and the deep integration of language acquisition, and reflects the positive promotion of technology-driven teaching innovation on language learning effects.

6. Conclusion

This paper proposes a multimodal intelligent discourse analysis model for high school English teaching. The model integrates multimodal data such as speech, text, facial expressions, and body movements, processes them through the Transformer architecture, models the dynamic context, and then uses deep reinforcement learning and knowledge distillation technology to achieve adaptive feedback. This method improves classroom interactivity and the stability and enthusiasm of students' emotional state in English learning and optimizes teaching results. However, this study still has shortcomings. The experiment does not cover the teaching differences in different regions, and the stability of the model in extremely complex scenarios needs to be improved. In the future, expanding the multimodal data type, further optimizing the generalization ability of the model, and exploring the application potential in different disciplines and teaching scenarios in order to provide stronger support for the development of intelligent education.

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

References

[1] Y. Yu, L. Han, X. Du, and J. Yu, "An oral English evaluation model using artificial intelligence method," *Mobile Information Systems*, vol. 2022, no. 1, p. 3998886, 2022. https://doi.org/10.1155/2022/3998886

[2] Q. Men, "Exploration of the application of artificial intelligence in high school English speaking teaching," *Communications in Humanities Research*, vol. 35, pp. 133-137, 2024.

[3] A. Moybeka, N. Syariatin, D. P. Tatipang, D. A. Mushthoza, N. Dewi, and S. Tineh, "Artificial intelligence and English classroom: the implications of AI toward EFL students' motivation," *Edumaspul: Jurnal Pendidikan*, vol. 7, no. 2, pp. 2444-2454, 2023. https://doi.org/10.33487/edumaspul.v7i2.6669

[4] U. Umar, "Advancements in English language teaching: Harnessing the power of artificial intelligence," *Foreign Language Instruction Probe*, vol. 3, no. 1, pp. 29-42, 2024.

[5] N. Sharma and H. Joshi, "Traditional Vs modern english language teaching methods: Study based on a survey," *MIER Journal of Educational Studies Trends and Practices*, vol. 14, no. 1, pp. 21-36, 2024.

[6] N. Djampulatova, "New methods of teaching English: Today and tomorrow," *International Journal of Artificial Intelligence*, vol. 1, no. 1, pp. 1089-1093, 2025.

[7] X. Li, "A personalized teaching system for college English based on big data and artificial intelligence," *Scalable Computing: Practice and Experience*, vol. 25, no. 6, pp. 5477–5485, 2024.

[8] A. P. B. Parraga *et al.*, "The impact of artificial intelligence on personalized learning in English language education," *Ciencia Latina Revista Científica Multidisciplinar*, vol. 9, no. 1, pp. 5500-5518, 2025.

[9] P. Binu, "Affordances and challenges of integrating artificial intelligence into English language education: A critical analysis," *English Scholarship Beyond Borders*, vol. 10, no. 1, pp. 34-51, 2024.

[10] R. Akbarani, "The use of artificial intelligence in English language teaching," *International Journal of English Learning and Applied Linguistics*, vol. 4, no. 1, pp. 14-23, 2023.

[11] S. Seemab, M. S. Zaigham, A. Bhatti, H. Khan, and U. A. Mughal, "Sentiment analysis of student feedback in higher education using natural language processing (NLP)," *Contemporary Journal of Social Science Review*, vol. 2, no. 04, pp. 1741-1750, 2024.

[12] S. S. Shahami and A. Tahriri, "English language teaching pages in focus: A multimodal discourse analysis," *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, vol. 43, no. 3, pp. 81-111, 2024.

[13] M. P. Pratama, R. Sampelolo, and H. Lura, "Revolutionizing education: Harnessing the power of artificial intelligence for personalized learning," *Klasikal: Journal of Education, Language Teaching and Science*, vol. 5, no. 2, pp. 350-357, 2023. https://doi.org/10.52208/klasikal.v5i2.877

[14] C. Mahmoud and J. Sørensen, "Artificial intelligence in personalized learning with a focus on current developments and future prospects," *Research and Advances in Education*, vol. 3, no. 8, pp. 25-31, 2024. https://doi.org/10.56397/RAE.2024.08.04

[15] A. Rudniy, "Artificial intelligence for automated scoring and feedback in chemistry courses," *Journal of Writing Analytics*, vol. 7, pp. 49–75, 2024. https://doi.org/10.37514/JWA-J.2024.7.1.04

[16] M. Hooda, C. Rana, O. Dahiya, A. Rizwan, and M. S. Hossain, "Artificial intelligence for assessment and feedback to enhance student success in higher education," *Mathematical Problems in Engineering*, vol. 2022, no. 1, p. 5215722, 2022. https://doi.org/10.1155/2022/5215722

[17] R. Duan and M. Li, "A study of high school english multimodal discourse teaching based on generative artificial intelligence," *The Development of Humanities and Social Sciences*, vol. 1, no. 1, pp. 1000077-1000077, 2025.

[18] X. Li, G. Han, B. Fang, and J. He, "Advancing the in-class dialogic quality: Developing an artificial intelligence-supported framework for classroom dialogue analysis," *The Asia-Pacific Education Researcher*, vol. 34, no. 1, pp. 495-509, 2025. https://doi.org/10.1007/s40299-024-00872-z

[19] Y. Yang and R. Hu, "The functions of context in discourse analysis," *International Journal of Linguistics, Literature and Culture*, vol. 8, no. 5, pp. 218-228, 08/15 2022. https://doi.org/10.21744/ijllc.v8n5.2172

[20] T. Ariwibowo, D. N. Hidayat, N. Husna, A. Alek, and A. Sufyan, "A discourse analysis of cohesion devices in students' writing of recount text," *Jurnal Pendidikan, Sains Sosial, dan Agama*, vol. 9, no. 1, pp. 22–32, 2023.

[21] P. S. R. Sihombing, H. Herman, and N. Saputra, "How to teach english conversation? An implementation of a multimodal discourse analysis through images," *English Review: Journal of English Education*, vol. 10, no. 2, pp. 431-438, 06/29 2022. https://doi.org/10.25134/erjee.v10i2.6244

[22] S. Triyono, W. Sahayu, and S. N. Fath, "Ecological discourse and environmental education in English textbooks: A multimodal eco-critical discourse analysis," *3L, Language, Linguistics, Literature*, vol. 29, no. 3, pp. 213-227, 2023. https://doi.org/10.17576/3L-2023-2903-15

[23] C. Li, N. Arumugam, G. Du, and X. Zhang, "Multimodal discourse analysis theory applied to senior high school english reading teaching," *Environment-Behaviour Proceedings Journal*, vol. 8, no. 26, pp. 101-106, 09/29 2023. https://doi.org/10.21834/e-bpj.v8i26.4957

[24] S. Herman, M. Ngongo, E. Fatmawati, and N. Saputra, "An analysis on multimodal discourse analysis towards English textbook used by students at school," *Journal of Namibian Studies*, vol. 33, pp. 613-624, 2023.

[25] Z. Liu, "Introducing a multimodal perspective to emotional variables in second language acquisition education: Systemic functional multimodal discourse analysis," *Frontiers in Psychology*, vol. 13, p. 1016441, 2022.

[26] C. Jiang and X. Mu, "A study on the application of discourse analysis theory in english reading teaching in senior high school," *GBP Proceedings Series*, vol. 3, pp. 144-153, 2025.

[27] H. Zhu and F. Amponstira, "Impact of artificial intelligence assisted approach on english teaching and learning," *Nimitmai Review Journal*, vol. 6, no. 1, pp. 38-55, 2023.

[28] P. Chea and Y. Xiao, "Artificial Intelligence in Higher Education: The Power and Damage of AI-assisted Tools on Academic English Reading Skills," *Journal of General Education and Humanities*, vol. 3, no. 3, pp. 287-306, 2024.

[29] S. S. Evenddy, "Investigating AI's automated feedback in english language learning," *Foreign Language Instruction Probe*, vol. 3, no. 1, pp. 76-87, 2024.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 6: 2536-2552, 2025
DOI: 10.55214/25768484.v9i6.8431