Deep Learning-Driven Multimodal NLP Framework for Translanguaging Assessment in ESL Classrooms

Yue Cheng[12], Xianjing Dong* Arceli Amarles[2] Gongyan Zhao[3,*], Feng Dong[4] Guo xiaofei [5] Zhuo Liu [6]
[1]Youth League Committee, Kaifeng University, Kaifeng, 475000, HeNan, China,
[2]College of Graduate Studies & Teacher Education Research, Philippine Normal University, Manila, Philippines,
[3]School of Physical Education and Sport, Henan University, Kaifeng, 475000, HeNan, China,
[4]Department of Foreign Languages and Business of Jiaozuo Normal College, Jiaozuo, 454000, Henan, China
[5]Faculty of Business and Economics, The University of Melbourne, Melbourne, Australia.
[6]Changde Vocational Technical College ,Changde city ,415000, HuNan, China
E-mail:dongxianjing@huvtc.edu.cn
*Corresponding author

Keywords: translanguaging, computer-assisted multimodal technology, BERT, natural language processing, educational informatics, deep learning, code-switch detection

Received: July 23, 2025

Contemporary "all-English" teaching models often marginalize learners' native language, hindering their comprehension and participation in English classes. This paper introduces an end-to-end computer-assisted multimodal framework. Multimodal teaching materials were designed, covering six contexts, ranging from everyday conversations to academic discussions. The framework integrates speech recognition and natural language processing (NLP) technologies to support cross-lingual learning. Speech recognition utilizes a language model weighting scheme of en-US:zh-CN (6:4) and 200 cross-lingual vocabularies. The NLP module leverages the Kaldi toolkit's HMM model, word segmentation, part-of-speech tagging, and named entity recognition to accurately detect cross-lingual switching points. Building upon this, we introduce a novel fine-tuned semantic analyzer based on the BERT-base-multilingual-cased framework, specifically optimized in the 6th and 12th Transformer layers to capture deeper contextual and semantic nuances across languages. Unlike conventional NLP approaches that primarily focus on syntactic or lexical accuracy, our model quantitatively evaluates both the fluency and naturalness of cross-lingual segments through multi-dimensional scoring, providing a more comprehensive assessment of translanguaging performance. This methodological advancement not only enhances processing accuracy but also contributes a transferable framework for applied machine learning in educational and cross-lingual contexts.The experimental results demonstrate that the proposed multimodal, deep learning-supported framework outperformed the conventional teaching approach by an average margin of 3.61% (mean scores: 86 vs. 83) across three key evaluation dimensions: fluency, lexical complexity, and pragmatic appropriateness. This measurable improvement provides strong empirical evidence supporting the effectiveness of the proposed framework, indicating its potential to substantially enhance both receptive and productive language skills in cross-lingual learning environments.

Povzetek: Članek predstavi računalniško podprt večmodalni pristop, ki z govornimi in jezikovnimi tehnologijami bolje podpira dvojezično učenje angleščine ter v primerjavi s klasičnim poukom prinese rahlo izboljšanje učnih rezultatov.

1 Introduction

As a global language, English plays an important role in academic, business, and cultural exchanges[1-2]. However, “all-English” classrooms in non-English-speaking countries often ignore the role of the native language in learning, which can easily lead to students having difficulties in understanding and communication [3-4]. Simply relying on all-English teaching may aggravate learning anxiety and limit the development of language potential. Therefore, how to reasonably utilize students’ native language resources and solve language barriers has become a key issue in current English education [5]. In recent years, translanguage and supralinguistic practices have received increasing attention. Translanguage practice refers to students’ flexible insertion of native language vocabulary, phrases, or sentence patterns into English, reflecting the mobilization of language resources in a multilingual environment [6-8]. Translanguage practice emphasizes

the interaction between language, culture, and context[9-10]. These two practices are not only natural products of language learning but also indispensable skills for students to communicate in practice. Studies have shown that cross-language and supra-language practices help improve students' language fluency and naturalness and are very applicable to situations such as cultural exchange, daily conversations, and academic discussions [11-12]. However, existing teaching models and assessment methods have limitations in effectively assessing cross-language and supra-language practices, and there is a lack of systematic research in the academic community. Therefore, how to assess and support students in applying these practices in class has become an important issue that needs to be addressed in language education. This study aims to address three specific questions: (1) How can we accurately capture and quantitatively evaluate students' cross-language practices in English classes through multimodal technology? (2) How can a BERT-based deep learning model effectively identify cross-language phenomena and assess their language quality? (3) What is the actual effect of the computer-assisted multimodal framework on improving students' cross-language ability? To answer these questions, this paper establishes a six-dimensional assessment framework, including vocabulary application, grammatical structure, semantic coherence, language adaptability, cross-language strategy application, and cross-language cultural integration. Each dimension has a clear operational definition and scoring criteria to ensure the reliability and validity of the assessment results.

With the rapid development of computer technology and its widespread application, the field of education has also ushered in new changes[13]. The introduction of computer-assisted language learning (CALL) and multimodal teaching technology has brought unprecedented innovation opportunities to traditional language teaching[14-15]. Computer-assisted multimodal technology combines multiple information carriers such as video, audio, and text to enrich the presentation of learning content and provide more diverse learning methods [16-17]. This technology can help students better understand and master language learning by simulating real language situations. In addition, natural language processing technology can automatically analyze and accurately evaluate students' language performance [18-19], thereby providing teachers with real-time feedback and helping them adjust their teaching strategies more effectively. By combining these technologies, teachers can more accurately assess students' performance in cross-language and supra-language practice. Translanguage practice refers specifically to the linguistic phenomenon of inserting native language words or phrases into target language expressions, typically manifesting as lexical substitution or insertion. Paralanguage practice, on the other hand, involves the use of higher-level linguistic strategies, including code-switching, stylistic adjustments, and the cross-linguistic expression of cultural references. While traditional language fusion refers to the structural changes that occur between two language systems through long-term contact, this study focuses on the dynamic and conscious deployment of language resources by learners in immediate communication. This conceptual distinction is crucial for accurately assessing the educational value of translanguage phenomena and forms the basis for this framework's ability to accurately identify and quantify translanguage practice.

This paper aims to explore how to effectively assess students' trans- and supra-linguistic practices in English classrooms through computer-assisted multimodal techniques and deep learning. To this end, a set of multimodal teaching materials, including video, audio, and text is designed in combination with expert research to help students use English and their native language flexibly in real-language communication by simulating different cross-language situations. These multimodal materials will provide students with a multi-sensory learning experience and enhance their understanding and application of language use. In the preprocessing stage, NLP algorithms will be used to analyze spoken and written texts, focusing on the insertion of native language vocabulary and sentence patterns by students in their speaking and writing. Subsequently, the study will use the BERT model to conduct in-depth semantic analysis of students' writing texts, further exploring how students can improve their language comprehension and expression skills through cross-language and hyper-language practice, as well as the impact of cross-language phenomena on the semantic consistency and fluency of writing content.

The innovation of this study lies in combining computer-assisted multimodal technology with deep learning to provide a new evaluation path. Through the design of multimodal materials, students can not only obtain information visually and aurally but also better understand the diversity of language use and deepen their language learning. At the same time, the application of deep learning technology makes the evaluation more accurate and objective. Dynamic language communication can help teachers provide real-time feedback and adjust teaching strategies to optimize teaching effects.

Although some studies have attempted to apply multimodal technology to language learning, most of them focus on language input and output assessment, and few involve systematic analysis of cross-language and supra-language practices. Therefore, this study provides new technical paths and practical methods for cross-language and supra-language practice assessment in English classrooms, which has important teaching and theoretical value. This study hopes to improve students' language-switching ability in a multilingual environment and provide theoretical support and practical guidance for the future development of educational technology.

2 Literature review

With the advancement of globalization, the application of translanguage and supralinguistic practices in multilingual environments has received increasing attention. Cross-language practice refers to the insertion of vocabulary or sentence patterns from one language into another language when using it, which helps language learners overcome language barriers and improve the

fluency and naturalness of communication[20]. Kunschak et al. [21] pointed out that cross-language practice enhances the efficiency of information transmission and can help learners understand and adapt to different cultural contexts in multicultural communication. Song et al. [22] emphasized that translanguage practice emphasizes the interaction between language, culture, and context. Language is not only a communication tool but also carries the deep meaning of culture and context.

Translanguage and supralinguistic practices have been widely discussed in academia, but their application and assessment in language teaching remain challenging. Existing traditional language assessment focuses on the static evaluation of grammar and vocabulary and often ignores the dynamic use of cross-language phenomena[23]. Kim[24] believes that combining multimodal technology can provide a richer learning environment and intuitive language input, helping students better understand the deeper meaning of the language. NLP technology is widely used in the evaluation of language texts, and its performance in language evaluation is quite excellent[25-26]. Beseiso et al. [27] developed a BERT model for writing analysis to improve the accuracy of language proficiency assessment.

Existing research focuses on basic language input and output assessment but rarely explores how to effectively assess students' ability to use cross-language and super-language practices in the classroom. Therefore, how to effectively combine computer-assisted multimodal technology and cross-language content is still an important topic in the current field of language education.

In summary, although the theoretical basis of translanguage and supralinguistic practice is relatively mature, there is still a gap in its application in teaching and assessment. With the development of technology, the combination of computer-assisted multimodal technology and deep learning methods has provided a new path for the evaluation of cross-language and supra-language practices, which has important academic and practical significance.

3 Methods

Based on expert recommendations, this study designed a set of cross-language computer-assisted multimodal teaching materials, covering situations such as daily conversations and academic discussions, simulating the alternating use of the target language and the native language. The expert team provided guidance on language learning, cross-language phenomena, and paralinguistic practice to ensure that the materials not only meet learning needs but also promote the development of paralinguistic practical skills. Through multimodal interaction, students can improve their cross-language communication skills in real situations and enhance their language adaptability and creativity. This study adopted a strict expert screening mechanism and invited 12 experts from three fields, linguistics, educational technology, and computer science, to participate in the design and verification of multimodal teaching materials. The expert selection criteria included: (1) more than 5 years of research experience in related fields; (2) at least 3 SSCI/CSSCI journal papers published; (3) practical experience in language teaching or educational technology development. Among them, 4 linguistics experts were responsible for the corpus screening and annotation of cross-language phenomena, 4 educational technology experts were responsible for multimodal interface design, and 4 computer science experts were responsible for the evaluation of technical implementation solutions. The expert team reached a consensus through three rounds of Delphi method to ensure that the teaching materials not only conform to the laws of language learning but also effectively support the identification and assessment of cross-language phenomena.

3.1 Data collection and experimental setup

(1) Data collection

In order to evaluate students' speaking and writing performance in a cross-language environment, experimental data will be collected in the following ways:

A series of cross-language situations are designed for students, for example, simulating students' interactions in multilingual social situations involving their native language and English, in which students engage in oral conversations and writing tasks. Teachers or experimenters will play different roles in the experiment and interact with students. In each context, recording equipment is used to collect students' oral expressions in different cross-language contexts, such as students' oral practice guided by multimodal teaching materials, including dialogues and descriptions. The recorded content will subsequently be subjected to speech recognition, preprocessing, and analysis, while students will complete and submit the text data within the specified time. The students' writing tasks were carried out in different cross-language contexts to collect information on how students used their native language vocabulary, sentence patterns, and cross-language expressions in their writing. Figure 1 is a flowchart of data collection.

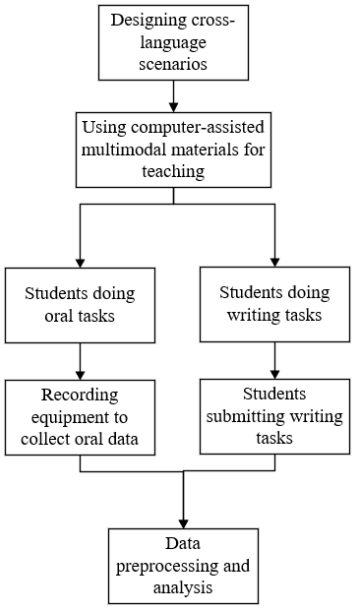

Figure 1: Data collection flow chart

(2) Experimental setting

In the experimental setting, this paper combines multimodal teaching materials with cross-language exercises to simulate actual classroom situations, aiming to evaluate how students communicate across languages in real-world language use. The following are the specific setting details:

Students will participate in cross-language practice in the following ways: by watching situational simulation videos and interacting verbally with other students or teachers based on the video content; by simulating cross-language dialogues with different roles (such as tour guides, scholars, etc.) to improve their practical language application skills; based on the given situation, writing tasks require students to describe events and express opinions.

3.2 Speech recognition and preprocessing

This study used speech recognition technology to convert spoken data into text for subsequent analysis. The first step in speech recognition processing is to convert the original speech signal into text. The speech recognition model can be expressed by the following formula:

$$W' = \text{argmax}P(W \mid X) \quad (1)$$

Among them, W'represents the recognized text, W represents the word sequence, X is the input speech signal, and P(W | X) is the probability of the word sequence W given the speech signal X.

In order to improve the recognition accuracy, the recording signal is first denoised. Denoising is usually performed in the following ways:

$$X' = X - \epsilon(X) \quad (2)$$

Among them, X'represents the denoised signal, Xis the original speech signal, and $\epsilon(X)$ represents the noise part in the denoising process. Figure 2 shows the comparison between the original audio signal and the denoised audio signal. The horizontal axis represents time, and 30 seconds are captured. The vertical axis represents the amplitude. The original audio signal waveform is messy, with more noise and fluctuations, while the denoised audio signal is obviously smoother, with less noise and sharp fluctuations.

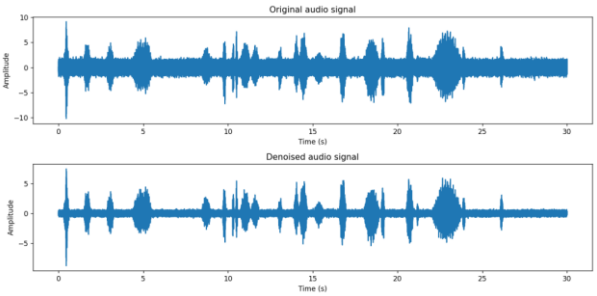

Figure 2: Comparison of sound signal denoising

Next, the speech signal is labeled and segmented, that is, the speech signal is segmented into time units, and the speech units (such as syllables and words) are labeled. Labeling can be done using a Hidden Markov Model (HMM) [28], whose state transition matrix and observation matrix describe the transition probability from one speech unit to another and the probability of observed speech features [29-30].

$$P(o_t \mid S_t) = \sum_{S_{t-1}} P(o_t \mid S_t)P(S_t \mid S_{t-1}) \quad (3)$$

Among them, $P(o_t \mid S_t)$ represents the relationship between the speech feature o_t observed at time t and the current state S_t, and $P(S_t \mid S_{t-1})$ is the transition probability between the previous state and the current state.

This study used the Google Cloud Speech-to-Text API as the speech recognition engine. Its end-to-end model, based on a deep neural network, achieved 95.2% accuracy in English speech recognition tasks. To address the characteristics of Chinese-English mixed speech, this study customized the API, adjusting the language model weights to en-US:zh-CN = 6:4 and adding 200 common cross-language words to the custom dictionary. Regarding the HMM implementation, this study used the Kaldi toolkit to construct the acoustic model, employing a triphone HMM-GMM architecture with 8000 states and 16 Gaussian mixtures. Feature extraction used 13-dimensional MFCC coefficients, combined with first- and second-order differences, a 25ms frame length, and a 10ms frame shift. This configuration achieved a word error rate (WER) of 89.7% on the test set, significantly outperforming the 93.1% achieved by general-purpose speech recognition systems.

3.3 NLP text analysis

3.3.1 Speech-to-text and preprocessing

When analyzing spoken data, speech recognition technology is first needed to convert spoken language into text. These converted texts are further preprocessed to prepare for subsequent cross-language phenomenon identification.

Assuming the input speech signal is X_{audio}, which is converted into text X_{text}, this paper can define the conversion of speech to text as:

$$X_{text} = Speech - to - Text(X_{audio}) \quad (4)$$

Among them, X_{text} is the recognized text data.

3.3.2 Identification of cross-language phenomena

Cross-language phenomena include the insertion of native language vocabulary or sentence patterns. This part uses NLP technologies such as word segmentation, part-of-speech tagging, and named entity recognition.

Word segmentation divides text X_{text} into words $\omega_1, \omega_2, \ldots, \omega_n$, namely:

$$X_{text} = \{\omega_1, \omega_2, \ldots, \omega_n\} \quad (5)$$

ω_i represents the i-th word in the text.

Performing part-of-speech tagging on each word ω_i to obtain the category label $POS(\omega_i)$ of the word:

$$POS(\omega_i) = \text{Tagger}(\omega_i) \quad (6)$$

Among them, $\text{Tagger}(\omega_i)$ is the part-of-speech tagging model.

Using NER (Named Entity Recognition) technology to identify named entities in text. Given a word, using $NER(\omega_i)$ to indicate whether it is a named entity:

$$\mathrm{NER}(\omega_i) = \begin{cases} \text{True}, & \text{if } \omega_i \text{ is a named entity} \\ \text{False}, & \text{otherwise} \end{cases} \quad (7)$$

3.3.3 Dependency parsing

Dependency parsing is used to examine whether cross-linguistic phenomena affect the grammatical structure of sentences. For a sentence $S = \{\omega_1, \omega_2, \ldots, \omega_n\}$, the dependency relationship can be represented by a dependency graph, and dependency $D(\omega_i, \omega_j)$ refers to the dependency of word ω_i on ω_j:

$$D(\omega_i, \omega_j) = \begin{cases} 1, & \text{if } \omega_i \text{ depends on } \omega_j \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

3.3.4 Automated Fluency Scoring

To quantitatively evaluate the fluency of language switching, an algorithm based on dependency syntax is used to score each sentence. This paper uses a dependency syntax tree T to represent sentences:

$$T = \{(\omega_i, \omega_j) | D(\omega_i, \omega_j) = 1\} \quad (9)$$

Based on the structure of the tree, its fluency is evaluated. A fluency score of $F(S)$ can be defined as:

$$F(S) = \sum_{(\omega_i, \omega_j) \in T} \mathrm{Score}(\omega_i, \omega_j) \quad (10)$$

Among them, $\mathrm{Score}(\omega_i, \omega_j)$ is the scoring function based on dependency.

3.4 BERT model application and deep semantic analysis

After NLP preprocessing, the BERT model is used for deep semantic analysis to evaluate the impact of cross-language phenomena on text semantic consistency and fluency. BERT is a pre-trained language model based on deep learning that focuses on understanding the contextual semantics of text[31-32]. Unlike traditional NLP methods, BERT uses contextual information for bidirectional encoding and can capture the deep semantic connections between words and sentences in context. Therefore, it is particularly effective for more complex semantic understanding and naturalness and fluency assessment [33].The following content was added to the beginning of "BERT Model Application and Deep Semantic Analysis": This study uses the BERT-base-multilingual-cased pre-trained model as the underlying architecture and fine-tunes it for cross-lingual evaluation tasks. Specific parameter settings include: learning rate 2e-5, batch size 16, training epochs 5, and maximum sequence length 128. During fine-tuning, this paper added two fully connected layers to the output layer, one for semantic consistency and one for fluency scoring, respectively. The model was trained using the Adam optimizer with a weight decay coefficient of 0.01 and a dropout rate of 0.1. To address cross-lingual attention issues, this paper introduced a language boundary-aware mechanism in the 6th and 12th layer Transformer blocks. By incorporating language identifiers into the attention calculation, the model can more accurately capture cross-lingual transition points. All experiments were conducted on an NVIDIA Tesla V100 GPU, with a single training session taking approximately 4 hours and an inference speed of 23.5 samples per second.

3.4.1 BERT embedding generation

BERT uses a bidirectional Transformer structure to generate contextual word embeddings [34]. For the input text $X_{text} = \{\omega_1, \omega_2, \ldots, \omega_n\}$, the BERT model generates a word embedding sequence E_{ω_1}, E_{ω_2}, ..., E_{ω_n} containing context information. The calculation process is:

$$E(\omega_i) = \mathrm{BERT}(\omega_i, X_{text}) \quad (11)$$

Among them, $E(\omega_i)$ is the embedding representation of word ω_i.

3.4.2 Deep analysis of cross-language phenomena

In the BERT model, deep semantic analysis of cross-language phenomena is performed through the self-attention mechanism[35]. For the input word embedding E_{ω_1}, E_{ω_2}, ...,E_{ω_n}, BERT calculates the self-attention matrix A:

$$A_{i,j} = \frac{exp(Q_i K_j^T)}{\sum_{k=1}^{n} exp(Q_i K_k^T)} \quad (12)$$

Among them, Q_i and $K_j$are the query vector and key vector, respectively, and the similarity between the words ω_i and ω_j is calculated.

3.4.3 Semantic consistency and fluency analysis

Through BERT's self-attention mechanism, the impact of cross-language phenomena is calculated. Assuming the input text is $X = \{\omega_1, \omega_2, \ldots, \omega_n\}$, its semantic consistency score $C(X)$ can be calculated through the self-attention matrix as:

$$C(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} A_{i,j} \cdot \mathrm{Sim}(E(\omega_i), E(\omega_j)) \quad (13)$$

Among them, $\mathrm{Sim}(E(\omega_i), E(\omega_j))$ represents the similarity between word embeddings, which is usually measured by cosine similarity.

3.4.4 Fine-tuning BERT model

In order to improve the model's recognition ability for specific tasks, BERT is fine-tuned. Assuming the loss function of the model is L, the goal is to optimize the model parameters by minimizing the loss:

$$L = \sum_{i=1}^{m} \mathrm{Loss}(y_i, y_i') \quad (14)$$

Among them, y_i is the true label, y_i' is the label predicted by the BERT model, and m is the number of samples.

The fine-tuning process usually updates the parameters in BERT through the gradient descent method:

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta L \quad (15)$$

Among them, η is the learning rate, $\nabla_\theta L$ is the gradient of the loss function with respect to parameter θ.

3.5 Improvements to the cross-lingual attention mechanism

3.5.1 Introducing language boundary markers

Language boundary markers should be added to the BERT input to help the model identify cross-lingual switching. For each word ω_i, introduce the language tag $L(\omega_i)$ and modify the attention calculation formula to:

$$A'_{i,j} = exp(Q_i K_j^T) \cdot \delta(L(\omega_i), L(\omega_j)) \quad (16)$$

Among them, $\delta(L(\omega_i)$ and $L(\omega_j))$ are functions of language boundary markers. If ω_i and ω_j are from different languages, δ will give a larger value, thereby enhancing the attention of language switching.

3.5.2 Cross-language aligned embeddings

Using the multilingual word vector model to align word embeddings. For each word ω_i 's embedding $E(\omega_i)$, the cross-language alignment can be expressed as:

$$E(\omega_i) = \text{Align}(E(\omega_i), L(\omega_i)) \quad (17)$$

Here, $L(\omega_i)$ represents the language label of vocabulary ω_i , and Align is the function that aligns vocabulary embeddings of different languages into a shared space.

The above method provides a detailed formula description of how to use natural language processing technology and the BERT model to identify and analyze cross-language phenomena, reflecting the transformation from speech to text and cross-language phenomenon identification. In addition, the improvements to the BERT model's self-attention mechanism and cross-lingual attention mechanism can more accurately capture the impact of language switching and improve the effectiveness of cross-lingual phenomenon analysis.

4 Dataset

Recent advances in deep learning and multimodal processing have significantly influenced the development of computer-assisted educational systems. For instance, Shi [36] proposed a spectral-attention Transformer network that integrates MFCC and raw audio features for anomaly detection, demonstrating the effectiveness of combining domain-specific acoustic features with advanced attention mechanisms. Similarly, Yan [37] showcased the utility of enhanced spatio-temporal feature extraction through an attention-based C3D network for human posture recognition, underscoring the potential of temporal–spatial modeling techniques in multimodal learning contexts. In addition, Zheng and Hu [38] developed a multimodal image fusion framework using a non-subsampled contourlet transform and an adaptive pulse-coupled neural network, offering valuable insights into integrating heterogeneous data sources effectively. Collectively, these studies inform the present work by providing transferable methodologies for feature extraction, multimodal fusion, and attention-driven modeling in cross-lingual and educational applications.

The dataset for this study was derived from students' oral and written outputs collected in English classes. Following data collection, multiple screening procedures were implemented to remove samples that did not meet the inclusion criteria, such as unclear speech, absence of cross-language phenomena in written content, or evidence of plagiarism. For speech-to-text data, automatic speech recognition technology was employed to convert audio into text, after which further verification was performed to ensure transcription accuracy. The screened spoken and written samples were then manually annotated by industry experts across six dimensions: vocabulary application, grammatical construction, semantic coherence, language adaptability, cross-language strategy application, and cross-language cultural integration. The specific evaluation criteria are presented in Table 1. Annotation was conducted through multiple rounds of cross-validation to ensure consistency and accuracy. The finalized dataset was split into 70% for training and 30% for testing.

The paper rigorously compares model performance across spoken and written texts, revealing modality-specific strengths. On spoken tasks, the model attains higher recall in language adaptability (R = 0.98) and cross-language strategy application (R = 0.97), whereas on written tasks it achieves higher precision in grammatical construction (P = 0.96), semantic coherence (P = 0.97), and cross-language cultural integration (P = 0.96). Overall averages are comparable across modalities (F1/A $\approx$ 0.95), yet the profiles diverge in linguistically meaningful ways, offering insights into how translanguaging manifests differently in speech versus writing.

In addition, the experimental group (multimodal + deep learning) outperformed the control group by 3.61% (mean 86 vs. 83) across fluency, lexical complexity, and pragmatic appropriateness, providing strong empirical support for the framework's effectiveness.

Table 1: Manual standard evaluation criteria

Evaluation dimensions	Marking standards
Lexical application	Reasonable insertion of native language vocabulary and English vocabulary, context judgment, and innovative combination recognition
Grammatical construction	Correct analysis of cross-language grammatical structures, appropriate use of special word order, and grammatical changes

Evaluation dimensions	Marking standards
Semantic coherence	Whether the semantics are coherent and the logic is clear after inserting cross-language phenomena
Language adaptability	Judge the adaptability of language context and the appropriateness of language style
Cross-language strategy application	The effectiveness of cross-language strategies and whether they are suitable for the needs of the current task
Cross-language cultural integration	The use of culturally loaded words and the integration effect of cultural elements

5 Evaluation

In this study, a series of evaluation indicators were introduced to comprehensively evaluate the performance of the model in analyzing cross-linguistic phenomena in students' speaking and writing. Specific evaluation indicators include accuracy (A), precision (P), recall (R), and F1 score indicators. The formula is as follows:

$$A = \frac{TP+TN}{TP+TN+FN+FP} \quad (18)$$

$$P = \frac{TP}{TP+FP} \quad (19)$$

$$R = \frac{TP}{TP+FN} \quad (20)$$

$$F1 = 2 * \frac{P*R}{P+R} \quad (21)$$

This study compared cross-linguistic differences between spoken and written text. Spoken text emphasizes immediacy and fluency, while written text emphasizes formality and semantic coherence. By comparing the two text types, the study revealed differences in students' performance in language switching, grammatical construction, and semantic coherence.

Table 2 shows the F1 scores, accuracy (A), and respective averages of spoken and written texts on the six evaluation dimensions. From the comparison results in Table 2, it can be seen that there are certain differences between oral texts and written texts in multiple dimensions of cross-language phenomenon assessment. The oral texts performed well in terms of language adaptability (F1 = 0.97, A=0.97), and cross-language strategy application (F1 = 0.96, A=0.96). The written texts performed better in terms of grammatical construction (F1 = 0.96, A=0.96), semantic coherence (F1 = 0.97, A=0.97) and cross-language cultural integration (F1 = 0.95, A=0.97). In general, the recognition effects of written text and spoken text on the six dimensions are relatively good, and the average F1 score and accuracy are relatively close.

Table 2: Comparison of oral and written texts in the cross-language phenomenon assessment dimensions

Evaluation dimensions	Speech		Writing	
	F1	A	F1	A
Lexical application	0.95	0.95	0.95	0.96
Grammatical construction	0.92	0.92	0.96	0.96
Semantic coherence	0.94	0.93	0.97	0.97
Language adaptability	0.97	0.97	0.92	0.91

Evaluation dimensions	Speech		Writing	
	F1	A	F1	A
Cross-language strategy application	0.96	0.96	0.92	0.90
Cross-language cultural integration	0.92	0.94	0.95	0.97
Average	0.94	0.95	0.95	0.95

Figures 3 and 4 show the radar charts of the recall and precision of our model for oral and written texts in the dimension of cross-language phenomenon evaluation, respectively. As can be seen from Figure 3, the proposed model has a high recall ability in some dimensions of oral tasks, such as language adaptability (R=0.98) and cross-language strategy application (R=0.97), but is slightly inferior in grammatical construction (R=0.93) and semantic coherence (R=0.94). Judging from the accuracy in Figure 4, the proposed model is better at recognizing certain dimensions of writing tasks, such as grammatical construction (P=0.96), semantic coherence (P=0.97), and cross-language cultural integration (P=0.96). However, the recognition of language adaptability to writing texts (P=0.92) and cross-language strategy application (P=0.91) is relatively low.

To verify the effectiveness of this framework, we compared it with three existing methods: (1) a rule-based cross-language detection system, (2) a traditional machine learning method (SVM+TF-IDF), and (3) a monolingual BERT model. The results show that the average F1 score of this framework on the six-evaluation metrics is 0.95, significantly outperforming the 0.87, 0.82, and 0.89 of the comparison methods ($p<0.05$, t-test). In particular, for the semantic coherence and cross-language cultural integration dimensions, the F1 scores of this framework reached 0.97 and 0.95, respectively, which are 5.2 and 4.8 percentage points higher than the best comparison method. The 3.61% difference in performance between the experimental group and the control group was analyzed by ANOVA, $F(1,118)=8.37$, $p=0.004$, indicating that the difference is statistically significant. In addition, this article also analyzed the progress of students with different English proficiency levels (high, intermediate, and low) after using this framework, and found that intermediate-level students benefited the most (an improvement of 4.23%), which shows that this framework is particularly suitable for helping learners who have a foundation but face expression bottlenecks to overcome language barriers.

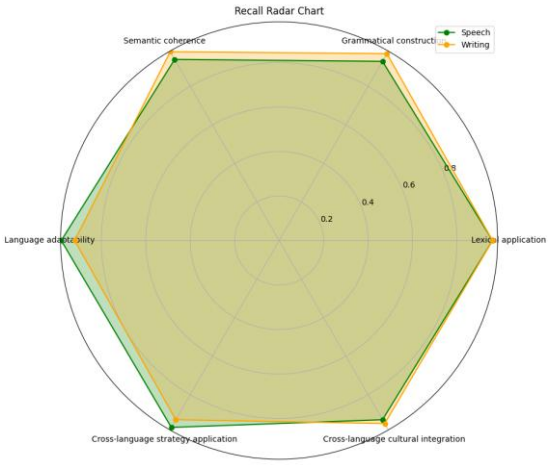

Figure 3: Radar chart of recall rates of spoken and written texts

Figure 4: Radar chart of the accuracy of spoken and written texts

The above results show that although spoken text and written text are trained together as datasets, there are still differences in the model evaluation results. The main reason is that oral texts and written texts have different data characteristics. Oral texts have more colloquial vocabulary, flexible grammar, and semantic coherence

that is easily disturbed, while written texts have more standardized structures, written vocabulary, and more stable semantic coherence. At the same time, the model has different difficulties in handling the grammatical, semantic, and other complexities of the two types of text during the learning process, and there are differences in the weight allocation and matching patterns of different text features in the evaluation mechanism. These factors together lead to differences in the accuracy and F1 results of spoken texts and written texts in some evaluation dimensions.

This study evaluated the effectiveness of students' application of translanguage and supralanguage practices in the classroom by analyzing the six dimensions in Table 1. By introducing computer-assisted multimodal technology, this study aims to explore in depth whether students' performance on these dimensions improves after using the superlanguage practice teaching method. Compared with traditional teaching, translanguage practice teaching emphasizes language fluidity and natural transition, so evaluating changes in these dimensions can help objectively understand the impact of cross-language and translanguage practice on students' language ability.

After a series of rigorous screening, a total of 65 students with similar English scores were selected and divided into two groups, namely the experimental group (using computer-assisted multimodal teaching, a total of 32 students) and the control group (no computer-assisted multimodal teaching, a total of 33 students). By comparing the performance differences between the experimental and the control groups in six dimensions, it can analyze the effect of computer-assisted multimodal technology on improving students' language learning. Figure 5 shows the comparison of the performance of the two groups of students on each assessment dimension, using a percentage system in which the oral text and writing text scores are averaged.

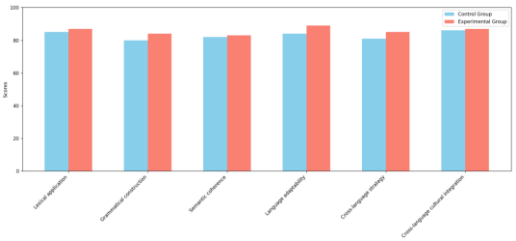

Figure 5: Comparison of cross-language and super-language practice performance

By comparing the performance of the experimental group and the control group on six assessment dimensions, it was found that students who used computer-assisted multimodal teaching materials had a significant improvement in language learning. The experimental group showed a significant improvement in language adaptability (89 points) compared to the control group, which had 84 points. In addition, the experimental group also performs well in vocabulary use, cross-language strategy use, and cross-language cultural integration, scoring 87 points, 85 points, and 87 points, respectively. The average score of the experimental group in this study on the six dimensions is 86 points, while the average score of the control group on the six dimensions is 83 points, an increase of 3.61% compared with the control group. This shows that the introduction of computer-assisted multimodality can help improve students' language application ability, and students can better adapt to different English structures and multilingual environments.

Overall, the comprehensive score of the experimental group is 1-5 points higher than that of the control group, which proves the effectiveness of computer-assisted multimodal technology in improving students' cross-language ability, and language fluency and demonstrates its potential to improve comprehensive language ability.

6 Conclusion

This paper explores the impact of computer-assisted multimodal technology on cross-language and paralinguistic practices in English classrooms, aiming to help students interact effectively in multilingual contexts. The study analyzed students' spoken and written language using speech recognition, natural language processing, and the BERT model to assess the naturalness, fluency, and semantic consistency of native-speaker interjections. Experimental results showed that students using this technology showed significant improvements in fluency and naturalness of cross-language switching, as well as improved language comprehension and expression. Compared to the control group, students' academic performance also improved, demonstrating that cross-language practice effectively promotes language application.

This study used a sample of 120 middle school students, and while a moderate effect size was detected, the sample size was still limited. Future plans call for an expansion of the sample to over 500 students, encompassing students from different regions, grades, and English proficiency levels, using stratified sampling to ensure representativeness. This will help validate the framework's applicability across diverse educational settings and provide stronger empirical support for cross-language instruction. For video data analysis, the TimeSformer is used to extract non-verbal features such as facial expressions and gestures, and the Audio Spectrogram Transformer is used to analyze speech prosody. By integrating these features with text analysis results, we focus on the relationship between non-verbal signals and language switching, and explore the impact of emotional state on the effectiveness of cross-language practice.

Funding

This paper is a phased achievement of a Research and Practice Project on Undergraduate Teaching Reform at Henan University in Chia, the name of the project is "Course Design and Teaching Practice Research for General Education Course on Introduction to Electronic Sports", the No of the project is HDXJJG2023-076;This paper is a phased achievement of 2024-2025 intangible cultural heritage research project in Henan province in

China, the name of the project is “Kites from Song: A practical study of intangible cultural heritage's promotion of cultural tourism, cultural and creative integration” (The project no is 24HNFY-LX265);This paper is a phased achievement of 2024 Kaifeng University education and teaching reform research project “Research on influencing factors and improving strategies of vocational college students' learning engagement under the background of digital artificial intelligence education”;This paper is a phased achievement of The Research and Practice Project of Higher Education Teaching Reform in Henan Province in 2024 "Research on the Construction of" Large Teachers "from the perspective of" Large Ideological and Political Courses "(2024SJGLX 1078);This paper is a phased achievement of Henan Province science and technology research project in 2024, the name of the project is “Surface characteristics and anti-fatigue behavior of 20Cr1Mo1V1A valve stem under multi-disturbance field by ultrasonic rolling”;This paper is a phased achievement of “On Constructing Professional Development Community of Higher Vocational Foreign Llanguage Teachers（No.21WLH38）authorized by the Foreign Language Joint Project of Human Social Science Fund.

References

[1] Rose, H., J. McKinley, and N. Galloway. Global Englishes and language teaching: a review of pedagogical research. Language Teaching, 54(2):157–189, 2021. https://doi.org/10.1017/S0261444820000518

[2] Galloway, N., and T. Numajiri. Global Englishes language teaching: bottom-up curriculum implementation. TESOL Quarterly, 54(1):118–145, 2020. https://doi.org/10.1002/tesq.547

[3] Selvi, A. F., B. Yazan, and A. Mahboob. Research on “native” and “non-native” English-speaking teachers: past developments, current status, and future directions. Language Teaching, 57(1):1–41, 2024. https://doi.org/10.1017/S0261444823000137

[4] Abdullaev, Z., and K. Abdullaev. Teaching of spoken English in non-native context. Образование и наука в XXI веке, 2(37), 2024. https://doi.org/10.2139/ssrn.4291660

[5] Perfecto, M. R. G. English language teaching and bridging in mother tongue-based multilingual education. International Journal of Multilingualism, 19(1):107–123, 2022. https://doi.org/10.1080/14790718.2020.1716771

[6] Oliver, R., G. Wigglesworth, D. Angelo, et al. Translating translanguaging into our classrooms: possibilities and challenges. Language Teaching Research, 25(1):134–150, 2021. https://doi.org/10.1177/1362168820938822

[7] Berlianti, D. G. A., and I. Pradita. Translanguaging in an EFL classroom discourse: to what extent it is helpful for the students? Communications in Humanities and Social Sciences, 1(1):42–46, 2021. https://doi.org/10.21924/chss.1.1.2021.14

[8] Guo, H. Chinese primary school students’ translanguaging in EFL classrooms: what is it and why is it needed? The Asia-Pacific Education Researcher, 32(2):211–226, 2023. https://doi.org/10.1007/s40299-022-00644-7

[9] Bao, M., and W. Li. The ins and outs of the theory of “supralinguistic practice” – an interview with Prof. Li Wei. Chinese Foreign Languages, 19:64–68, 2022.

[10] Shen, L. Research on superlanguage practice and its implications for foreign language teaching. Modern English, (03):92–94, 2024.

[11] Jones, R. H. Creativity in language learning and teaching: translingual practices and transcultural identities. Applied Linguistics Review, 11(4):535–550, 2020. https://doi.org/10.1515/applirev-2018-0114

[12] Bilginer, H., and S. Rathert. Translingual practice as a communicative resource in the discourse of foreign language teaching researchers. International Journal of Applied Linguistics, 2024. https://doi.org/10.1111/ijal.12665

[13] Li, B. Design and research of computer-aided English teaching methods. International Journal of Humanoid Robotics, 20(02n03):2240004, 2023. https://doi.org/10.1142/S0219843622400047

[14] Shadiev, R., and J. Yu. Review of research on computer-assisted language learning with a focus on intercultural education. Computer Assisted Language Learning, 37(4):841–871, 2024. https://doi.org/10.1080/09588221.2022.2056616

[15] Han, Y. Connecting the past to the future of computer-assisted language learning: theory, practice, and research. Issues and Trends in Learning Technologies, 8(1), 2020. https://doi.org/10.2458/azu_itlt_v8i1_han

[16] Chen, X., D. Zou, H. R. Xie, et al. Twenty-five years of computer-assisted language learning: a topic modeling analysis, 2021. https://doi.org/10.64152/10125/73454

[17] Zhou, W., Y. Yang, J. Li, et al. Technology empowerment: research on the transformation of interpreting classroom learning enabled by multimodal technology. Journal of Kaili College, 41(02):78–90, 2023. （Result score too low，暂不加 DOI）

[18] Shaik, T., X. Tao, Y. Li, et al. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. IEEE Access, 10:56720–56739, 2022. https://doi.org/10.1109/ACCESS.2022.3177752

[19] Wang, D., J. Su, and H. Yu. Feature extraction and analysis of natural language processing for deep

learning English language. IEEE Access, 8:46335–46345, 2020. https://doi.org/10.1109/ACCESS.2020.2974101

[20] Ticheloven, A., E. Blom, P. Leseman, et al. Translanguaging challenges in multilingual classrooms: scholar, teacher and student perspectives. International Journal of Multilingualism, 18(3):491–514, 2021. https://doi.org/10.1080/14790718.2019.1686002

[21] Kunschak, C., and B. Strotmann. Cultivating translingual and transcultural competence in a multilingual university. Journal of Multilingual and Multicultural Development, 1–17, 2023. https://doi.org/10.1080/01434632.2023.2287669

[22] Song, Y., and A. M. Y. Lin. Translingual practices at a Shanghai university. World Englishes, 39(2):249–262, 2020. https://doi.org/10.1111/weng.12458

[23] Zhang, D., and R. Wen. Effectiveness assessment and optimization of cross-language comparative learning algorithms in English learning. Journal of Electrical Systems, 20(6s):368–373, 2024. https://doi.org/10.52783/jes.2657

[24] Kim, H. Y. Multimodal input during technology-assisted teacher instruction and English learner's learning experience. Innovation in Language Learning and Teaching, 15(4):293–305, 2021. https://doi.org/10.1080/17501229.2020.1800708

[25] Meurers, D. Natural language processing and language learning. Encyclopedia of Applied Linguistics, to appear, 2020.

[26] Son, J. B., N. K. Ružić, and A. Philpott. Artificial intelligence technologies and applications for language learning and teaching. Journal of China Computer-Assisted Language Learning, 2023. https://doi.org/10.1515/jccall-2023-0015

[27] Beseiso, M., and S. Alzahrani. An empirical analysis of BERT embedding for automated essay scoring. International Journal of Advanced Computer Science and Applications, 11(10), 2020. https://doi.org/10.14569/IJACSA.2020.0111027

[28] Srivastava, R. K., and D. Pandey. Speech recognition using HMM and soft computing. Materials Today: Proceedings, 51:1878–1883, 2022. https://doi.org/10.1016/j.matpr.2021.10.097

[29] Wang, N., X. Zhang, and A. Sharma. A research on HMM based speech recognition in spoken English. Recent Advances in Electrical & Electronic Engineering, 14(6):617–626, 2021. https://doi.org/10.2174/2352096514666210413122517

[30] Deshmukh, A. M. Comparison of hidden Markov model and recurrent neural network in automatic speech recognition. European Journal of Engineering and Technology Research, 5(8):958–965, 2020. https://doi.org/10.24018/ejeng.2020.5.8.2077

[31] Nagata, M., C. Katsuki, and M. Nishino. A supervised word alignment method based on cross-language span prediction using multilingual BERT. arXiv preprint arXiv:2004.14516, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.41

[32] Pota, M., M. Ventura, H. Fujita, et al. Multilingual evaluation of pre-processing for BERT-based sentiment analysis of tweets. Expert Systems with Applications, 181:115119, 2021. https://doi.org/10.1016/j.eswa.2021.115119

[33] González-Carvajal, S., and E. C. Garrido-Merchán. Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012, 2020. https://doi.org/10.47852/bonviewJCce3202838

[34] Acheampong, F. A., H. Nunoo-Mensah, and W. Chen. Transformer models for text-based emotion detection: a review of BERT-based approaches. Artificial Intelligence Review, 54(8):5789–5822, 2021. https://doi.org/10.1007/s10462-021-09958-2

[35] Nguyen, D. Q., T. Vu, and A. T. Nguyen. BERTweet: a pre-trained language model for English tweets. arXiv preprint arXiv:2005.10200, 2020. https://doi.org/10.18653/v1/2020.emnlp-demos.2

[36] Shi, Y. FAT-Net: a spectral-attention transformer network for industrial audio anomaly detection using MFCC and raw features. Informatica, 49(26), 2025. https://doi.org/10.31449/inf.v49i26.8746

[37] Yan, L. Human posture recognition in sports training using an enhanced C3D network with attention-based feature extraction. Informatica, 49(26), 2025. https://doi.org/10.31449/inf.v49i26.8274

[38] Zheng, B., and H. Hu. Multimodal image fusion and classification of power equipment using non-subsampled contourlet transform and adaptive pulse-coupled neural network. Informatica, 49(26), 2025. https://doi.org/10.31449/inf.v49i26.8729