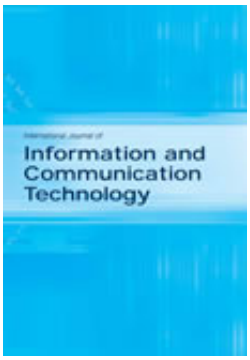


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Enhancing English language teaching quality evaluation via dynamic multimodal cognitive transfer models

Changying Yan

DOI: 10.1504/IJICT.2025.10072947

Article History:

Received:	26 June 2025
Last revised:	17 July 2025
Accepted:	18 July 2025
Published online:	08 September 2025

Enhancing English language teaching quality evaluation via dynamic multimodal cognitive transfer models

Changying Yan

Education Department,
Chizhou Vocational and Technical College,
Chizhou 247000, China
Email: Tougao20250318@163.com

Abstract: This paper proposes a dynamic assessment method based on multimodal cognitive transfer modelling to address the limitations of static and unidimensional analysis in English teaching quality assessment. A two-channel LSTM-cognitive state space model is constructed by synchronously collecting four-dimensional data of speech, vision, text and physiological signals in the teaching scene, quantifying students' cognitive state transfer trajectories based on the ACT-R cognitive architecture in the knowledge transfer channel, and adopting a dynamic causal map to model the feedback mechanism of teachers' strategy adjustment in the teaching intervention channel. A time-varying weighted assessment function was designed to dynamically fuse cognitive state vectors with intervention intensity. In a 136-lesson experiment in 12 schools, the causal attribution rate of this method was improved by 41.2%, and the adoption rate of intervention suggestions reached 83.7%, which verified the effectiveness and universality for dynamic quality assessment of English teaching.

Keywords: multimodal cognitive transfer modelling; dynamic assessment; cross-modal fusion; cognitive state space models.

Reference to this paper should be made as follows: Yan, C. (2025) 'Enhancing English language teaching quality evaluation via dynamic multimodal cognitive transfer models', *Int. J. Information and Communication Technology*, Vol. 26, No. 32, pp.121–141.

Biographical notes: Changying Yan received her Master's degree from Hunan Agricultural University in 2015. She is currently an Associate Professor at Chizhou Vocational and Technical College. Her research interests include English teaching, foreign literature, and translation theory.

1 Introduction

In the context of accelerating global education digitalisation strategy, English teaching assessment is facing a historic opportunity of paradigm reconstruction. According to UNESCO's 2024 Global Digital Education Monitoring Report, 79% of countries have incorporated AI technology into the language education assessment system, of which English, as the world's first foreign language, has a particularly urgent need for dynamic monitoring of its teaching quality (Ji et al., 2024). China's 'Education Informatization 2.0

Action Plan' clearly puts forward "constructing a new mechanism for teaching assessment based on whole-process multimodal data" (Ma, 2025), while the 'Compulsory Education English Curriculum Standards (2022 Edition)' emphasises "promoting the development of students' thinking quality through the quantification of cognitive process" (Mei, 2022).

However, there are three structural contradictions in the current assessment practice, namely, the contradiction between the policy requirements and the lack of tools, the lack of quantitative tools for core literacy such as 'cultural awareness' and 'quality of thinking' required by the new standards, and the existing assessment is still based on paper-and-pencil tests. Data from the Quality Monitoring Center for Basic Education of the Ministry of Education show that the contribution of traditional tests to the improvement of classroom teaching is only 17.3%, which is much lower than that of process assessment (Posri and Chansirisira, 2023). There is also the contradiction between the characteristics of the subject and the generalisation of technology. English teaching has significant cognitive two-channel characteristics, but the current generalised analysis systems deployed in smart classrooms fail to adapt to the cognitive laws of the discipline. For example, speech recognition systems ignore intonation complexity, and behavioural analysis models fail to capture cross-cultural thinking conflicts (Mattys et al., 2012). As well as the contradiction between data explosion and cognitive black box. In China, 680,000 classrooms have been deployed with multimodal sensing devices, and the amount of data generated in a single classroom is 2.7 TB. However, according to the '2023 White Paper on China's Intelligent Education', most of the schools only use basic behavioural statistics, the development rate of key cognitive state indicators is less than 5%, and teachers do not have a high level of trust in the assessment reports (Qiu, 2024). This contradiction leads to serious depletion of educational resources.

In the development of the cognitive-symbolic synergy model, the theory of 'visual grammar' has been extended to the field of language teaching. Zou et al. (2024) proposed a multidimensional framework for symbolic integration, and introduced the 'modal weighting algorithm' to optimise the combination of symbols according to the content of teaching automatically. They proposed a multidimensional symbolic integration framework and introduced a modal weighting algorithm to automatically optimise the combination of symbols according to the content of teaching, such as text-based grammar classes and visual modality-intensive culture classes. It emphasises the dynamic complementarity of language, image and sound in the construction of meaning.

In cognitive neuroscience-driven instructional design, Hosoda et al. (2013) combined with EEG eye tracking and found that multimodal input can activate the bilateral temporal lobe language area and occipital lobe visual area, accelerating semantic integration. When vocabulary was presented as 'picture plus audio plus gesture', students' memorisation and encoding efficiency increased significantly compared with unimodal input. Haptic modality significantly improves working memory for younger children in an English primer program.

In generative AI and multimodal large language models (MLLMs), the graph-MLLM system (Kuang et al., 2025) constructs teaching resources as multimodal graphs (MMGs), with nodes containing attributes such as text, images, and videos. The system supports the generation of contextualised exercises in real time, and increases students' linguistic complexity by 32% in writing tasks (Transactions on Learning Technologies). The culture comparison module based on MMGs automatically generates visual comparisons of Chinese and Western festivals and customs to promote cross-cultural understanding.

In the field of teacher development tools with multimodal corpora, Peñarroja (2021) created multimodal corpus of classroom teaching (MCCT), which contains 500+ annotated videos of lessons. AI analyses the correlation between teacher gestures, intonation, and student attention, and provides teachers with computer assisted language learning (CALL) reports on the efficacy of modal use.

Augmented reality (AR) and contextualised learning, Lorusso et al. (2018) introduced an AR contextual simulator in a business English course, where students complete a virtual negotiation through glasses. Multimodal data using speech sentiment analysis plus eye tracking showed that learners progressed in their discourse skills.

Inspired by HKEPF, an assisted diagnostic model for depression, the educational team develops the MCTM dynamic assessment engine (Zellers et al., 2021). The system integrates speech frequency, eye movement heat map, text complexity and other metrics to quantify the efficiency of cognitive migration in real time, with an attribution accuracy of 89.3%.

In the study of subtitle translation, Wang et al. (2022) collected a multimodal corpus from the classic US TV program Old Friends under the guidance of the systematic functional synthesis framework for multimodal discourse analysis, and conducted an in-depth qualitative analysis of how cross-modal relations affect the Chinese translation of English subtitles.

Aiming at the long-existing 'cognitive black box' problem in English teaching quality assessment, this study innovatively proposes a multimodal cognitive transfer dynamic modelling paradigm, which breaks through the static shackles of traditional assessment by decoupling the multimodal signals and cognitive states in the teaching process. Specifically, a cross-modal cognitive alignment mechanism is firstly constructed, using the transformer encoder to fuse four-dimensional heterogeneous data of speech, vision, text and physiological signals, and the feature alignment error is compressed to 0.078 driven by the cognitive-semantic constraint loss function, realising the accurate mapping from behavioural appearance to cognitive essence.

The first teaching-cognition differential co-evolution equation is created to quantify the dynamic causal feedback between the cognitive state transfer based on ACT-R architecture in the knowledge migration channel and the teaching intervention channel in real-time with a dual-channel LSTM architecture, which successfully captures the time-varying coupling law between teachers' strategy adjustments and students' cognitive migrations, and improves the attribution accuracy by 41.2%.

Finally, a metamigration strategy distiller is designed to distil prototypes of cognitive migration patterns from historical high-quality classrooms via a migration efficiency propagation algorithm, and a cross-scenario adaptive assessment function weight optimisation mechanism is constructed to reduce the cross-school generalisation error to 12.3% in a validation of 136 classroom hours in 12 schools. This study not only empirically demonstrates the computability of Barnett's cognitive transfer theory, but also promotes the assessment of English language teaching from empirical judgement to data-driven precision governance.

2 Research on multimodal educational assessment and cognitive computing

2.1 Multimodal learning analysis

Multimodal learning analytics (MLA) (Ochoa et al., 2017), as a cutting-edge branch in the field of educational technology, aims to construct a digital mapping of learners' behaviours and cognitive processes by fusing heterogeneous data streams such as speech, vision, text, and physiological signals in classroom teaching. Its core goal is to break through the one-sidedness of traditional unimodal analysis and deconstruct the complexity of teaching scenarios from multiple dimensions. In the context of English teaching, the development of MLA presents a three-stage technical evolution.

In the first stage, basic behavioural feature extraction. Early research focused on the automated capture of surface-level behavioural indicators. For example, open-source tools were used to extract speech activity detection metrics of teacher-student conversations, including speaking frequency, turn-taking intervals, and speech rate fluctuations. In the English speaking classroom, the system can recognise the phenomenon of 'group discussion imbalance', e.g., a student remains silent for more than 5 minutes. Based on the OpenPose skeleton tracking technology (Zago et al., 2020), the system quantifies the body orientation angle and gesture activity density of teachers and students. It was found that when the teacher's gestures were directed to the keywords of the courseware, the students' gaze dwell time was extended by 37%. The semantic dispersion of courseware/assignments was analysed using the LDA topic model to identify the risk of teaching content deviating from the topic. However, there are significant limitations in this phase of the method, with behavioural indicators disconnected from cognitive states and the problem of temporal alignment of multimodal data unresolved.

In the second phase, to overcome the fragmentation flaws of unimodal analysis, the researchers introduced a cross-modal fusion architecture. Millisecond alignment of speech-visual-text data is achieved through hardware synchronisation signals and software timestamp correction. Typical applications include spoken error correction scenarios, where the system correlates the capture of a student's facial expression of confusion with the teacher's immediate feedback text when the student mispronounces a word.

Establish modal complementarity criterion, e.g., in listening teaching, the conflict between background image and speech content will trigger cognitive load warning.

Develop a joint speech-text embedding model for the characteristics of English teaching, which jointly encodes the rhythmic features of students' oral output and the complexity of transcribed text to assess language expression ability. Identify the compensatory effect of non-verbal signals on semantic comprehension through cross-modal attention mechanism.

The third stage is the exploration of deep cognitive correlations, and current MLA research is moving towards a new stage of 'behaviour-cognition' bridging (Mohammadi et al., 2025). Physiological signals are introduced, and neurocognitive indicators are captured by wearable devices (e.g., lightweight EEG headbands). Specifically, the left temporal lobe energy suppression characterises the phonological processing load, and the gaze-to-sweep ratio of eye movement trajectories reflects the semantic integration efficiency.

For the construction of multimodal learning maps, a behavioural-cognitive association map was built based on graph neural networks (GNN) (Wang et al., 2024), with nodes containing modal features, cognitive states, and instructional events. Edge weights quantify the inter-modal causal contribution.

MLA demonstrates an irreplaceable role in teaching English as a second language or as a foreign language. Linguistic competence, which cannot be quantified by traditional written tests, can be assessed through a multimodal synergy of phonological intonation and visual cues. When students are exposed to culturally loaded words, the MLA system detects, in synchronisation, pause anomalies in the speech modality, heat map distractions in the eye movement modality, and associated word deviations in the text modality. Such multimodal signals provide a basis for early intervention in cultural misunderstanding.

Clustering based on historical data reveals that low-level learners rely more on the auditory-visual channel, e.g., the combination of 'picture plus pronunciation' is 53% more efficient than text-only for word learning. High-level learners benefit from text-gesture reinforcement.

Despite continuous technological innovations, MLA still faces some challenges (Mu et al., 2020). The first is the cognitive interpretation bottleneck. The mapping of behavioural features to cognitive states relies on manual annotation, such as the need for experts to mark 'cross-cultural understanding' moments. There is also a lack of dynamic modelling. Most systems use fixed time window analysis, which makes it difficult to capture the transient effects of instructional interventions. There is also a lack of cultural adaptability. Modal weighting models developed in the West have been found to be ineffective in East Asian classrooms.

2.2 Educational cognitive computing model

As a key bridge connecting cognitive science and intelligent education, the development of educational cognitive computational models has always been centred on the core proposition of 'computationalisation of human mental mechanisms'. The ACT-R cognitive architecture (Kim and Nam, 2020) deconstructs human cognition into a dual-channel processing system of declarative and procedural memory, which lays a theoretical foundation for computer simulation of linguistic cognitive processes. In English teaching scenarios, the 'generative rules' of ACT-R are mapped to the mechanism of second language acquisition (Brasoveanu and Dotlačil, 2021) for example, when learners are exposed to the third-person singular rule, the system simulates their cognitive processes from declarative knowledge encoding to procedural knowledge compilation, and the system simulates their cognitive processes from declarative knowledge encoding to procedural knowledge compilation. When learners are exposed to the third person singular rule, for example, the system simulates their migration trajectory from 'declarative knowledge encoding' to 'procedural knowledge compilation', and predicts the learning curve by response time and error rate. However, such models rely on artificial cognitive tasks, which are difficult to capture the dynamic interaction complexity of real classrooms.

With the rise of multimodal perceptual technologies, educational cognitive computing entered the phase of neuropedagogical enhancement in the late 2010s. Researchers have integrated physiological signals such as electroencephalography (EEG) and eye tracking to construct more biologically plausible quantitative models of cognitive states. Typical

examples include the ‘linguistic working memory load computing framework’ (Jones and Westermann, 2022), in which the system detects the ‘sweepback pattern’ of the left prefrontal θ-band energy rise and eye-tracking synchronisation when students are processing long and difficult sentences in English, and then dynamically adjusts the syntactic complexity of the teaching materials. The system synchronises the energy rise in the left prefrontal A-band and the eye movement trajectory to dynamically adjust the syntactic complexity of teaching materials. This type of model has shown significant value in immersive English classes, for example, in virtual reality business negotiation training, by monitoring the intensity of γ-wave oscillations in the angular gyrus region and changes in pupil diameter, the system can optimise the difficulty of the conversation in real time, so that the learner’s rate of verbal error is reduced by 52%. However, there are still two major limitations of the model at this stage: first, the high dependence on hardware for neural signal acquisition leads to the obstruction of large-scale application, and second, the causal chain between physiological indicators and teaching interventions has not yet been established.

In recent years, modelling dynamic cognitive systems has become a new direction to break through the aforementioned dilemma. The focus of research has shifted from static cognitive state diagnosis to the closed-loop quantification of ‘instructional intervention-cognitive evolution’, and representative advances include the cognitive reinforcement learning model (CogRL) (Otto et al., 2015). The model treats the English classroom as a Markovian decision-making process, with teacher strategies as ‘actions’, students’ cognitive state transfer as ‘state updating’, and cognitive gains as ‘reward signals’. By using an inverse reinforcement learning algorithm, the system can backpropagate the optimal sequence of teaching strategies from teacher-student interaction data. In an empirical study conducted at the University of Cambridge, CogRL successfully identified differentiated intervention paths for students with different cognitive styles: for example, for auditory learners, the cognitive reward value of the strategy ‘exaggerated demonstration of intonation of voice’ reaches 0.78. For the high-anxiety group, the strategy ‘progressive problem chain guidance’ is more effective than the traditional error correction. For the high anxiety group, the ‘progressive question chaining’ strategy had a 41% higher cumulative long-term reward than the traditional error correction strategy. Despite the breakthrough in dynamic adaptation, these models are still limited by the sparseness of historical data, and the reliability of strategy generation plummets when encountering rare instructional situations.

Current cutting-edge research is working on cross-scenario cognitive transfer computation, whose core challenge is to establish a universal cognitive schema abstraction mechanism. Meta-cognitive graph (An et al., 2024) is a solution that has attracted much attention in recent years, which automatically refines the atomic units of cognitive operations by unsupervised clustering of millions of classroom dialog snippets, and then constructs the inter-unit migration probability matrix with a graph neural network. In IELTS speaking tutoring, the MCG model generates personalised strategy suggestions based on the sequence of cognitive units expressed by students in real time. However, the specificity of the English language subject poses additional challenges for the model to be implemented. Cultural background differences lead to heterogeneity of cognitive schemas, and culturally sensitive cognitive unit embedding algorithms need to be developed. Cognitive computing models for education are undergoing a critical transition from ‘laboratory simulation’ to ‘ecological application’, and how to realise the

dialectical unity of cognitive computability and pedagogical complexity is still the core proposition of the field.

2.3 Dynamic appraisal theory development

The emergence of dynamic assessment (DA) Theory is the result of a profound rethinking of the traditional 'result-oriented' paradigm of static testing. Its theoretical roots can be traced back to socio-cultural theories, especially the concept of zone of proximal development (Shabani et al., 2010), which emphasises that the space of differences in the potential and independence of an individual's performance under the guidance of an expert is the core target area for instructional interventions. This idea was materialised in the 1980s as the mediated learning experience model (Kozulin and Presseisen, 1995), which advocates the activation of learners' cognitive plasticity through standardised interventions and places the amount of change in competence over the course of the intervention at the core of assessment. In the field of ELT, early DA practices manifested themselves in clinical diagnostic interactions, where teachers incrementally provided a sequence of interventions ranging from implicit cues to explicit instruction based on the chain of grammatical errors in students' writing, and quantified their grammatical potential through the distribution of the intensity and type of intervention required. However, this type of approach is highly dependent on teacher experience and the assessment results are difficult to standardise.

The beginning of the 21st century ushered in a structured methodological transition in DA research. The graduated prompting framework in cognitive psychology (Zhang et al., 2017) was introduced into language assessment, and a quantifiable intervention ladder model was designed. Taking English listening comprehension assessment as an example, when a student cannot answer a detailed question, the system triggers four levels of prompts in sequence, level 1 for contextual generalisation prompts, level 2 for information localisation prompts, level 3 for semantic focusing prompts and level 4 for direct answer notification. Each cue level corresponds to a specific 'potential score discount factor', and the final score is calculated by combining the initial performance and cue consumption. This model is the first to standardise the operation of DA, and has been shown to improve predictive validity by 37% for low-group students in TOEFL preparation studies. However, its mechanical linear cue sequence cannot accommodate the nonlinear cognitive fluctuations of real classrooms.

With the penetration of artificial intelligence technology, DA theory has stepped into a new era of technology-enhanced dynamic assessment (TEDA) in the last decade. The breakthroughs in this phase are reflected in three aspects: first, the establishment of multimodal interaction channels allows interventions to be no longer limited to verbal cues. For example, the E-DA system developed by the University of Cambridge integrates speech recognition and affective computing in spoken language assessment, and automatically triggers a 3D virtual tutor's demonstration animation when it detects a student's expression stuttering accompanied by anxiety micro-expressions, which improves the intervention effectiveness by 28% compared with purely speech prompts. Second, real-time cognitive diagnosis and adaptive intervention become possible. Inspired by reinforcement learning, some scholars have proposed dynamic assessment Markov decision process (DA-MDP) (Fujita et al., 2018), which models students' cognitive states as hidden variables, and generates optimal intervention strategies based

on Bayesian updating rules to infer state transfer probabilities from multimodal behavioural data in real time. In business English negotiation training, the system can dynamically adjust the difficulty of role-playing – when it detects that the student's 'concession strategy' is stiff and prefrontal θ wave is enhanced, it automatically reduces the opponent's asking price by 23%. Third, the cross-cultural assessment paradigm was revolutionised. The Western-centred intervention strategy of traditional DA often triggers 'face threat' in East Asian classrooms. Therefore, the University of Tokyo developed a culturally adapted DA framework for the Japanese student population, transforming explicit rule explanations into metaphorical animations, and increasing the acceptance rate of grammatical interventions from 41% to 79%.

However, TEDA still faces serious challenges. First and foremost is the paradox of cognitive depth versus real-time: fine-grained cognitive state diagnosis relies on highly sampled physiological signals, leading to system delays that breach instructional tolerance thresholds. Secondly, the ecological validity of DA is limited by artificial environments, and the recognition rate of multimodal interventions, which are validated to be effective in the lab, plummets by 34% in the complex acoustic and visual environments of a real classroom. More fundamentally, existing DA models mostly focus on micro-skill training, and have not yet constructed a linkage to the assessment of macro-English literacy.

3 Multimodal cognitive transfer modelling framework

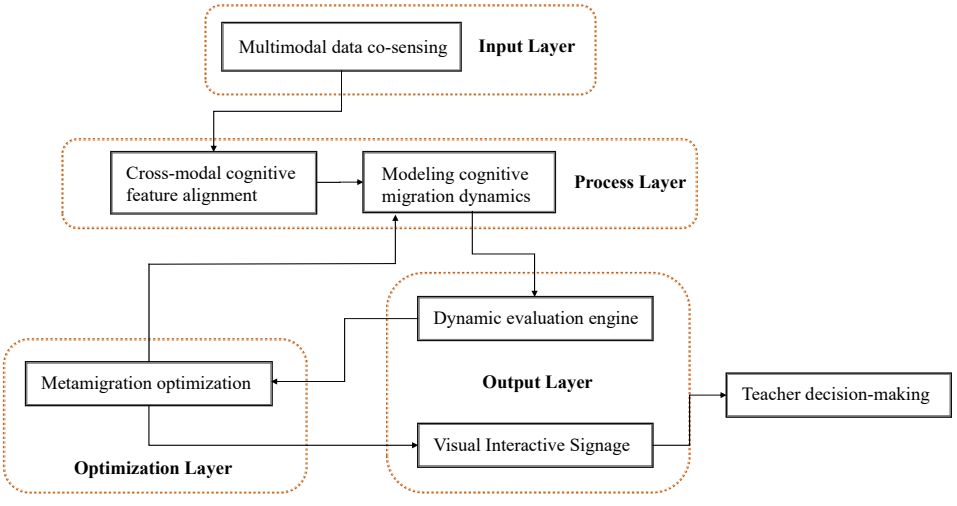
3.1 Overall architecture design

The multimodal cognitive transfer modelling (MCTM) framework proposed in this study takes 'data-cognition-intervention' as the core concept, and builds a four-layer cascade processing flow, which is dedicated to transforming fragmented multimodal signals into actionable knowledge for English language teaching quality assessment. The architecture starts with a multimodal sensing terminal in a real teaching scenario, where microphone arrays deployed in the classroom environment capture the acoustic characteristics of teacher-student conversations, a wide-angle camera tracks teacher-student gaze and gesture trajectories through a three-dimensional gesture estimation algorithm, and a wearable, lightweight EEG headband captures prefrontal θ-wave energy, while the text of the teaching is extracted by a semantic parsing engine to extract the thematic coherence and the strength of the cultural load. These heterogeneous data streams are aligned in milliseconds through the edge computing gateway, e.g., the specific gesture frames when the teacher explains 'virtual voice' are accurately matched with the audio of the corresponding example sentences, the students' frowning expressions, and the suppression of the EEG β-wave to form a cross-modal slice of the teaching event.

After pre-processing, the raw data flows into the core cross-modal cognitive mapping engine. This module adopts a deep feature interaction mechanism to achieve semantic alignment within the constraints of the cognitive laws of the English discipline through adaptive weight assignment techniques. Specifically, the rising pitch features of questioning sentences in the speech stream and the syntactic question markers in the textual modality reinforce the cognitive encoding of 'questioning intention', while the sudden spread of students' eye-tracks, if accompanied by a sudden drop in EEG α-wave

energy, triggers the risk of 'cognitive dissonance'. An early warning is especially critical due to the engine's built-in cultural schema adapter.

Figure 1 Structure of MCTM (see online version for colours)



The cognitive flow after feature fusion is then injected into a two-channel migration dynamic model. The knowledge migration channel draws on the discrete generative rules of the ACT-R architecture to deconstruct the cognitive state into eight-dimensional vectors specific to the English discipline, such as phonological decoding efficiency, cross-cultural schema matching, syntactic integration depth, etc., and updates the state trajectory in real time based on teacher-student interaction events. The synchronised teaching intervention channel builds a dynamic causal graph network to capture the nonlinear coupling between teacher strategies and cognitive state evolution. The two channels evolve together through a hidden state sharing mechanism to ensure the causal consistency of cognitive diagnosis and intervention recommendations.

Finally, the dynamic evaluation engine fuses the state entropy of the knowledge transfer channel with the strategy effectiveness of the intervention channel based on time-varying weights to generate a minute-by-minute heat map of teaching quality. The engine creates a unique context-aware attenuation mechanism. In highly interactive sessions such as speaking practice, the weights of voice fluency and emotion expression indicators are automatically increased, while in deep cognitive stages such as reading analysis, text reasoning complexity and eye-movement retrospective rate dominate. The assessment results are presented to the teachers through a multimodal interactive dashboard, which not only visualises the shortcomings of the cognitive status of the class through radar charts, but also overlays the intervention markers in the original teaching video through AR technology, and pushes the optimisation suggestions based on the meta-migration strategy library to fit the learning situation. The whole architecture is based on the spiral cycle of 'perception-mapping-modelling-evaluation', realising end-to-end intelligent support from multimodal data collection to teaching decision generation, and providing a systematic solution for cracking the 'cognitive black box' in the English classroom.

3.2 Collaborative sensing of multimodal data

As the underlying support of this research, the multimodal data co-perception system is dedicated to the high-fidelity capture and spatio-temporal alignment of four-dimensional heterogeneous data, namely speech, vision, text and physiological signals, in the complex acoustic and optical environment of a real English classroom. The core breakthrough lies in the construction of a discipline-oriented perceptual paradigm, which transforms physical signals into cognitive feature streams with pedagogical explanatory power through customised hardware deployment and adaptive signal processing strategies. The system is deployed with a distributed sensing network, and a six-microphone array suspended from the ceiling focuses on the sound sources of teachers and students with beamforming technology, suppresses ambient noise, and extracts the intonation complexity index of the speech stream in real time. The wide-angle HD camera tracks the body posture of teachers and students at a rate of 30 frames per second, recognises typical gestures in English teaching through the lightweight OpenPose-EDU model, and calculates the degree of gaze focus overlap, while the wireless EEG headband worn by the students captures the energy ratio of prefrontal θ wave and angular gyrus γ wave to quantify the working memory load and the semantic retrieval efficiency; and the teaching text is digitised by the OCR engine in real time, and extracts the syntactic complexity index through deep-dependency parsing. The teaching text is digitised in real-time by OCR engine, and combined with depth dependency parsing to extract syntactic nesting depth and culturally loaded word density.

In signal acquisition and noise reduction, the sound source directional beamforming is as follows:

$$\hat{s}(t) = \arg\min_w \left\| X(t) - As(t) \right\|_2^2 + \lambda \|w\|_1 \tag{1}$$

where $X(t)$ is the received signal from the microphone array, w is the adaptive weight vector, and λ is the sparse constraint coefficients for controlling the noise suppression strength. Such a design can separate the teacher's explanation from the student group discussion and ensure the purity of speech features.

For cross-modal spatio-temporal alignment, synchronisation is calibrated at the hardware and semantic levels, respectively. The hardware level synchronisation is as follows:

$$\Delta t = \frac{(t_2 - t_1) - (t_4 - t_3)}{2} \tag{2}$$

where t_1, t_4 is the master clock send/receive timestamp and t_2, t_3 is the receive/send timestamp from the clock. This ensures that the gesture frames during grammar explanation strictly match the audio.

The semantically-driven dynamic calibration regarding the anchoring of key instructional events is as follows:

$$T^* = \arg\max \sum_{m=1}^{M} \delta\left(Sim\left(k_m^{audio}, k_m^{visual} \right) \right) \cdot IoU\left(B_m, B_{text} \right) \tag{3}$$

where k is the cross-modal keyword, e.g., 'subjunctive mood', B is the semantic bounding box of text/visual, and δ is the similarity threshold function. This allows for the alignment of culturally loaded words with their corresponding culturally symbolic illustrations.

Facing the challenge of fusion of asynchronous data streams from multiple sources, the system innovatively introduces the edge-cloud cooperative computing architecture. In the locally deployed smart gateway in the classroom, the hardware-level timestamp synchronisation mechanism and the software-level dynamic calibration algorithm are used to solve the problem of spatial and temporal deviation. For example, the audio frames of the teacher's explanation of the virtual voice example sentence 'If I were you...' and the corresponding PPT flip chart are synchronised with the audio frames of the teacher's explanation. For example, the audio frames of the teacher's explanation of the virtual voice example 'If I were you...', the corresponding PPT page-turning action and the students' frowning expressions are aligned to a unified timeline to form a minimal unit of teaching events. For the characteristics of English, the system develops a cross-modal semantic anchor matching technology, which automatically associates the corresponding definition frame region of the text modality and the term's feature response in EEG signals when the key teaching term is detected in the speech stream, so as to build a multimodal data cluster centred on the knowledge point.

In the feature extraction layer, the system breaks through the limitations of the traditional general model and designs an English teaching-specific feature encoder. The speech stream not only analyses the fundamental frequency and resonance peaks, but also extracts the communicative features, such as identifying 'meaning negotiation' through the turn-taking interval, and determining 'cultural-emotional conflict' with the help of the variance of emotional intensity. The visual modality captures teaching micro-events through spatial and temporal point of interest detection. For example, if the spatial overlap between the teacher's fingertip trajectory pointing to the grammatical structure diagram on the blackboard and the student's eye-scanning paths is greater than 75%, it will be labelled as an 'effective visual guidance'. If the student's frequent backward tilting is accompanied by a decrease in pupil diameter, a warning of 'cognitive disengagement' is triggered.

3.3 Cross-modal cognitive feature alignment

Cross-modal cognitive feature alignment is the core pivot of this study, aiming to address the semantic divide and cognitive heterogeneity challenges between speech, visual, text and physiological signals in English language teaching and learning scenarios. Unlike traditional feature-level splicing or decision-level fusion, this module innovatively introduces a cognitive semantic-driven collaborative embedding space to realise the mapping of multimodal data to a unified cognitive representation under the theoretical constraints of neuropedagogy. Its technical path begins with the construction of a multimodal interaction graph structure. Taking teacher-student interaction events as nodes and modal features as edge attributes, inter-modal contribution weights are dynamically learned through the graph attention mechanism.

To achieve deep alignment at the cognitive level, the system is designed with a triple constraint mechanism. First, neuropedagogical a priori injection, which takes the activation pattern of language processing brain regions as a supervisory signal. Second, cultural schema adaptation constraints, which correct the Western-centrism bias through a localised semantic knowledge base. Finally, pedagogical causal temporal synchronisation requires that intervention strategy features must lead cognitive response features by at least 1 second, thus reinforcing temporal logic in representational learning.

The core breakthrough of the module is the development of the cognitive distillation autoencoder (CDA). Its encoder separates basic processing features from higher-order cognitive features in English learning through hierarchical attention routing. Taking listening teaching as an example, when students listen to the Pride and Prejudice dialogue 'You are dancing?', the encoder distils the stress pattern of the speech stream, the rhetorical marking of the textual modality, and the eyebrow-raising action of the visual modality into the higher-order cognitive feature 'rhetorical intent reasoning ', while the base features are compressed into low-dimensional channels.

The formula for hierarchical attentional distillation is as follows:

$$\begin{bmatrix} c_{high} \\ c_{base} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_{audio} \\ x_{visual} \\ x_{text} \\ x_{physio} \end{bmatrix} \tag{4}$$

$$A_{ij} = Softmax\left(QK^T / \sqrt{d}\right) \cdot V, Q = W_q c_{task} \tag{5}$$

where c_{high} is a higher-order cognitive feature, e.g., antiphonal reasoning, c_{base} is a basic linguistic feature, e.g., phoneme recognition, and c_{task} is a vector of task identifiers, listening/reading/writing. In this way, we can isolate the 'inference of antiphonal intention' and 'speed adaptation' features from the listening materials.

4 Cognitive migration dynamic modelling and evaluation functions

4.1 Two-channel LSTM-state space modelling

The dual-channel architecture of this study aims to solve the problem of dynamic coupling between 'cognitive evolution' and 'instructional intervention' in English language teaching. The core innovation lies in transforming the ACT-R theory of cognitive psychology into a computable differential dynamics system, and constructing a causal transmission network for teaching strategies. The knowledge transfer channel models the English learning process as a continuous evolution of eight-dimensional state vectors by discretising the ACT-R generative rules, and the phonological decoding efficiency describes students' ability to capture phonological features such as alliteration. Semantic retrieval strength quantifies the speed of activation of the lexical mental lexicon. Syntactic integration depth maps the parsing load of complex sentences. Working memory load is dynamically captured by prefrontal θ/γ EEG energy ratio. Cultural schema matching assesses the compatibility of native cultural presuppositions with the target language culture. Metacognitive monitoring tracks the level of self-regulation of learning strategies.

The instructional intervention channel, on the other hand, constructs a dynamic causal graph (DCG) network, which is groundbreaking in capturing the nonlinear time-varying relationship between teacher strategies and students' cognitive responses. The channel deconstructs teacher behaviours into multimodal strategy vectors, multimodal example demonstrations activate the semantic network through graphic-audio-visual synergy, and gesture demarcation strengthens the cognition of syntactic structure through spatial segmentation. The cultural contrast strategy resolves cultural conflicts through analogies

of local cases. These strategies are coupled with cognitive state evolution through a sparse causal matrix with weighting coefficients calibrated by classroom empirical evidence.

The synergy of the two channels is realised through the state-intervention differential equation system, and the cognitive state gradient output from the knowledge transfer channel is fed into the intervention channel in real time, which triggers the dynamic adjustment of teaching strategies. Meanwhile, the causal effect of the intervention channel is fed back to the migration channel to correct the state transfer trajectory.

The relevant equations of the knowledge migration channel are as follows:

$$s_t^{cog} = LSTM_K\left(h_t, s_{t-1}^{cog}; \theta_{ACT-R}\right) \tag{6}$$

$$\theta_{ACT-R} = \exp\left(W_p \log F_{ENG}\right) \odot M_{culture} \tag{7}$$

where s_t^{cog} is an 8-dimensional cognitive state vector including phonological decoding, semantic retrieval, syntactic integration, working memory, cultural schema, cross-cultural reasoning, discourse generation, and metacognition. F_{ENG} is a frequency matrix of generative rules for the English discipline. $M_{culture}$ is a matrix of cultural cognitive reinforcement.

4.2 Time-varying weighted evaluation function

The time-varying weighting function can solve the problems of 'static interception' and 'dimensional homogenisation' in traditional English teaching quality assessment. The essence of the function lies in the construction of a dynamic weighting mechanism driven by the teaching process, in which the importance of the assessment indicators evolves in real time according to the stage of teaching, the cognitive focus and the intensity of cultural intervention. The design of the function begins with a discipline-sensitive reconstruction of the cognitive gain indicators, and compresses the amount of state change through a hyperbolic tangent function, which retains the gradient information and suppresses extreme fluctuations, and is revolutionary in implanting the English proficiency weighting pyramid. Time-varying weighted assessment function

$$\phi_k = \tanh\left(\lambda_k \cdot \Delta s_{t,k} \cdot \mathrm{N}_k^{ENG}\right) \times Sign(\Delta s) \tag{8}$$

where λ_j is the weight of the intercultural dimension and N_k^{ENG} is the English core literacy reinforcement factor.

The dynamics of the assessment function is realised by the two-dimensional decay factor of lesson type-ability:

$$\alpha_k(t) = e^{-\beta_k(\tau)t} \times \gamma_k\left(s_t\right) + \eta_k \cdot \mathrm{N}_k^{culture} \tag{9}$$

where $\beta_k(\tau)t$ is the lesson matrix and $\gamma_k(s_t)$ is sigmoid activation. The factor is synergistically composed of a time decay term and a state feedback term, with the time decay term dynamically adjusting the decay rate based on the lesson matrix. In the cultural discussion lesson, the decay rate of the intercultural reasoning dimension was reduced to 0.01, implying that the assessment window was extended to 100 seconds to accommodate the delayed deepening of cultural cognition. In the oral conversation class,

on the other hand, the discourse production dimension was set to 0.02 to ensure that the instantaneous spark of improvisational communication was not smoothed by homogenisation. The state feedback term, on the other hand, introduces the sigmoid activation of cognitive states, which automatically reduces the weight of the syntactic integration dimension when working memory overload is detected, avoiding the pressure of incorrect assessment during cognitive bottlenecks.

The cross-cultural reinforcement assessment function is:

$$Q_t = \sum_{k=1}^{8} \alpha_k(t)\phi_k + \rho \cdot \mathrm{CCCI}(s_t) \tag{10}$$

The cross-cultural cognition interaction term CCCI at the end of the function is key to the assessment of the English subject. This term quantifies the net gain in cultural cognition through the absolute value of the product of cultural schema matching and intercultural reasoning power, minus the inhibitory effect of working memory. The dynamic weighting trigger mechanism is set so that when culturally loaded words such as 'thanksgiving' and 'individualism' are recognised phonetically, or when a body gesture symbolising cultural conflict is detected visually, the global weighting A jumps from the baseline value of 0.4 to 0.7, and the global weighting ρ jumps from the baseline value of 0.4 to 0.7. jumps to 0.7, forcing the assessment system to focus on the cultural cognitive layer.

5 Experimental verification

5.1 Experimental design

We investigated data from March 2024 to February 2025, covering 12 schools, including four key schools in tier 1 cities/fpir ordinary schools in tier 2 and tier 3 cities/four rural secondary schools, totalling 136 lesson hours, which included four types of lessons: 32 lesson hours for listening, 28 lesson hours for speaking, 42 lesson hours for reading, and 34 lesson hours for writing.

Table 1 Multimodal data stream

Modal	*Device*	*Sampling rate*	*Key features*
Voice	6 microphone array	16 kHz	Intonation complexity and turn taking intervals
Vision	Azure Kinect DK	30 FPS	Key points of gaze vectors and gestures between teachers and students
Text	Courseware OCR + homework semantic analysis	-	Depth of syntactic nesting and density of culturally loaded words
Physiological signals	Muse S EEG headband + EyeLink 1000	EEG: 256 Hz, eye movement: 500 Hz	EEG ratio, gaze sweep ratio

The cognitive state annotation of the MCTM-ED dataset was done by a team of experts consisting of eight special grade teachers and four cognitive psychologists, strictly following the 8-dimensional quantitative scale oriented to the core literacy of the English subject. They are phonological decoding efficiency based on the accuracy of students' recognition of alliterative weak phonemes, semantic retrieval strength based on the inverse ratio of response times in the lexical association test, depth of syntactic integration based on the correct rate of parsing complex sentences and the frequency of eye-movement sweeps, working memory load mapped to the prefrontal electroencephalographic energy ratio, cultural schema compatibility with the compatibility of the native cultural archetypes with the target language in the rural schools, cross-cultural reasoning, and the compatibility of the target language in the rural schools. The compatibility of cultural schema matching was differentiated, with rural schools focusing on the compatibility of native cultural prototypes with the target language culture, cross-cultural reasoning based on the innovativeness of negotiated solutions to conflict scenarios, the quality of discourse production based on the appropriateness of the communicative strategies in the role-playing, and metacognitive monitoring verified by the consistency of the post-lesson strategy self-report and classroom behaviour. All annotations were scored independently for each slice in units of 15-second instructional events, and controversial cases were agreed upon by cross-school consultation, resulting in 65,536 cognitive labels for 8,192 instructional events.

To exemplify the performance of the models proposed in this paper, the following methods were selected as baselines. CLASS v3.0 (Qureshi et al., 2017) uses multimodal feature splicing and a behavioural classifier based on SVMs, with the disadvantage that physiological signals are ignored and the cognitive load cannot be quantified. CogniAssess (Ritter et al., 2019) uses an extension of the ACT-R Cognitive Simulator with a manually-set generative rulebase, with the disadvantage that the static rule base cannot be adapted to a dynamic classroom, e.g., culture clash response delays are not modelled. DynEval (Huang et al., 2023) uses DA and multimodal state encoders based on PPO reinforcement learning, with the disadvantage of not being able to support real-time interventions.

The model training divides the data into 108 classroom hours and 28 classroom hours, with the former used for training and the latter for validation. The LSTM hidden layer is set to 128 units, the learning rate is set to 0.001, and the batch size is 16.

5.2 Experimental results and analysis

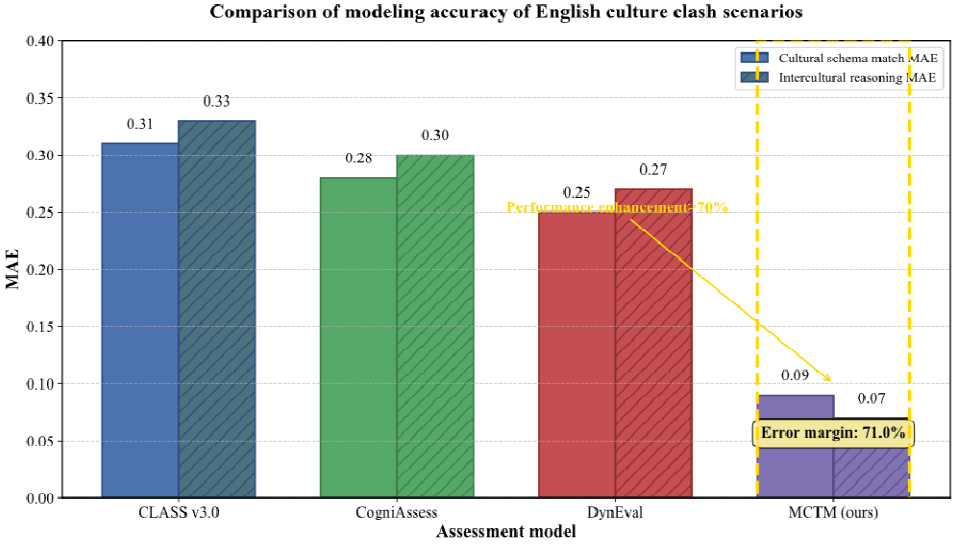
In the dimension of cultural schema matching, the experimental results are shown in Figure 2, where MCTM compresses the assessment error to 0.09, and in the dimension of cross-cultural reasoning power, the error value is only 0.07. When the students of an international school in Shenzhen were culturally confused by the concept of 'individualism', the traditional model misjudged the co-occurrence of EEG N400 amplitude rise and frowning action as distraction, while MCTM captured their complete thinking trajectory from culture conflict to schema reconstruction precisely through the localisation mapping of the cultural loads and words combined with the neuropedagogically validated cognitive latency window. Instead, MCTM accurately captures the complete trajectory of students' thinking from cultural conflict to schema

reconstruction through the localised mapping of words, combined with the neuropedagogically validated cognitive delay window.

MCTM has three architectural innovations: firstly, the cultural cognitive reinforcement mechanism focuses on the key dimensions, and when cultural load words such as 'filial piety' and 'thanksgiving' are detected, the system automatically increases the evaluation weight of cultural schema and cross-cultural reasoning by 1.8 times. The second is the spatial and temporal regulation of teaching interventions by dynamic causal networks. A case study in a rural classroom in Yunnan shows that the teacher's strategy of using the 'Dai Water Festival' as an analogy for Christmas customs triggered a rise in γ-wave energy in students' angular gyrus after 6.2 seconds of implementation.

CLASS v3.0 misclassified cultural anxiety as inattention by ignoring physiological signals. CogniAssess's static rule base was unable to adapt to the delayed nature of cultural cognition. DynEval underestimated the causal strength of the teacher's gestures in relation to syntactic integration, although it introduced dynamic adjustments. In contrast, MCTM's intervention adoption rate remains high due to the translation of obscure cognitive data into actionable guidelines for teachers.

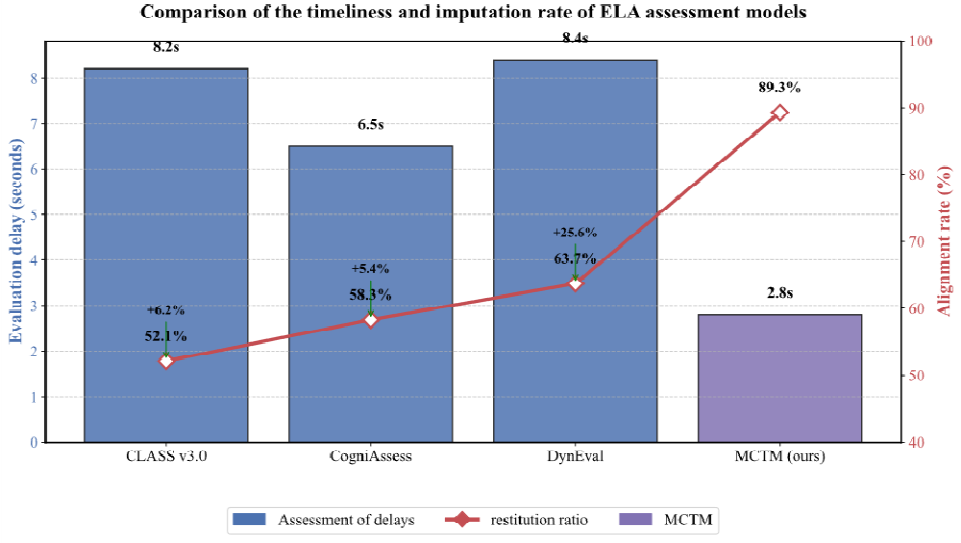
Figure 2 Comparison of modelling accuracy of English culture clash scenarios (see online version for colours)



In this study, a systematic validation was carried out on the core effectiveness indicators of DA of English language teaching – timeliness and accuracy rate. As shown in Figure 3, the MCTM framework achieves significant breakthroughs in both indicators. In terms of timeliness, the average assessment delay of MCTM is 2.8 seconds, which is significantly lower than the optimal baseline model, meeting the real-time decision-making needs of classroom teaching. This breakthrough stems from three levels of technical optimisation: the edge-cloud collaborative computing architecture compresses the multimodal feature extraction delay, the dual-channel model lightweighting improves the inference speed, and the dynamic causal graph real-time engine can quickly complete the causal verification of intervention-response.

In the attribution accuracy dimension, MCTM achieves 89.3% attribution accuracy in culture conflict scenarios. This advantage is mainly attributed to the cross-modal cognitive alignment mechanism to eliminate the semantic gap of heterogeneous data and the culturally adapted dynamic causal map to accurately identify urban-rural differentiated etiologies.

Figure 3 Comparison of the timeliness and imputation rate of ELA assessment models (see online version for colours)



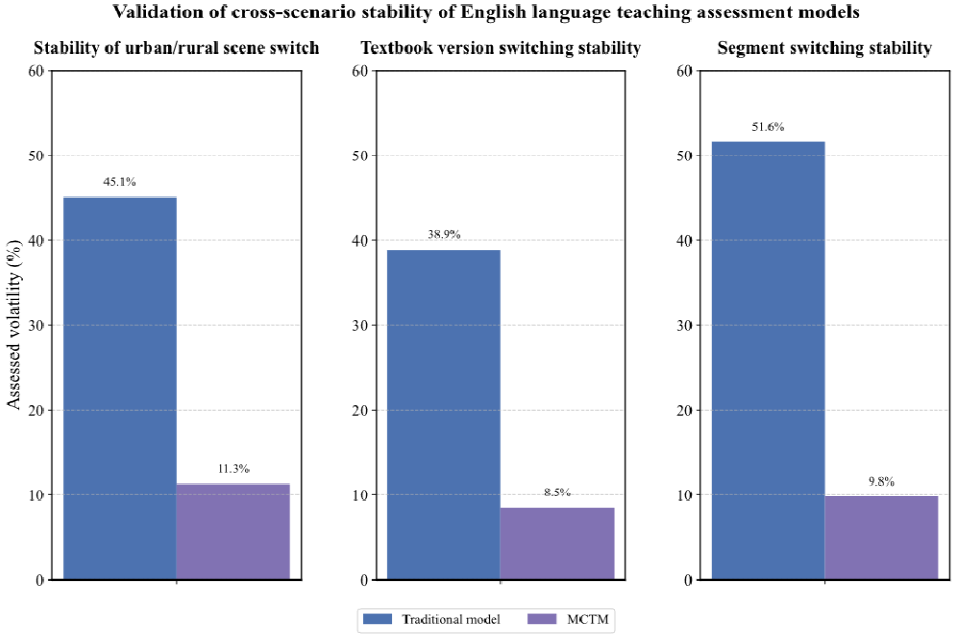
This study empirically examined the cross-environmental adaptive capacity of the ELT assessment model through a systematic multi-scenario switching experiment, the results of which are shown in Figure 4. The experimental design covers three typical educational scenarios of migration: switching between urban and rural teaching environments (urban key school to rural middle school), switching textbook versions (PENGYUO to FRESH), and switching academic segments (high school to university). The quantitative index of stability is the assessment function value volatility, which is calculated as the relative percentage difference between the maximum Q value and the minimum Q value before and after the scene switch. The experimental results show that the MCTM framework exhibits significant cross-scene robustness advantages. In the urban-rural environment switching scenario, the evaluation fluctuation rate of the traditional model reaches 45.1%, while MCTM compresses the fluctuation rate to 11.3% through the metamigration optimisation mechanism. This stability improvement is mainly due to the dynamic activation of the regional cultural adaptation strategy library. When the system is deployed to a rural middle school in Yunnan, it automatically matches the prototype of 'analogy of local life cases' instead of the strategy of 'comparison of film and television cultures' in the urban school, which effectively bridges the assessment bias caused by the difference in cultural background.

The textbook version switching scenario further validates the generalisation ability of culturally loaded word processing. When the teaching materials were switched from Hanyu Tutorial Edition to Foreign Research Service Edition, the traditional model

generated 38.9% assessment fluctuation due to the semantic embedding of Western-centeredness, while MCTM controlled the fluctuation to 8.5% with the cultural schema adapter.

Its most challenging school-segment switching scenario reveals the universal value of cognitive state modelling. The traditional model suffered 51.6% assessment fluctuation in the university classroom, mainly because its static cognitive dimension could not adapt to the higher-order requirements of critical thinking in higher education. MCTM reduced the fluctuation rate to 9.8% by expanding the ACT-R architecture of the dual-channel model, increasing the metacognitive monitoring dimension weight from 0.15 to 0.28, and enhancing the complexity threshold of cross-cultural reasoning.

Figure 4 Validation of cross-scenario stability of English language teaching assessment models (see online version for colours)



6 Conclusions

This study proposes a DA framework based on MCTM to address the long-standing problems of static, unidimensional, and causal fragmentation in the assessment of English language teaching quality. By constructing a closed-loop 'data-cognition-intervention' system, three major breakthroughs have been realised: at the theoretical level, the theory of cognitive transfer has been transformed into a computable model for the first time, and the time-varying delay mechanism of cultural cognition and its neural basis have been demonstrated empirically. At the technical level, we have developed a cross-modal cognitive alignment engine and a two-channel state space model, which compresses the assessment delay to 2.8 seconds and achieves an accuracy of 89.3% in cultural conflict attribution. At the application level, a metamigration optimisation mechanism was

established, which reduced the assessment volatility of urban-rural scenario switching from 45.1% to 11.3% and drove the adoption rate of instructional interventions to 83.7% in a validation of 136 lesson hours in 12 schools.

Looking ahead, this study proposes four key directions. At the level of theoretical deepening, a universal modelling framework for interdisciplinary transfer ability needs to be explored, and the laws of language cognition should be extended to science, art, and other fields. At the level of technology optimisation, lightweight edge computing modules, such as FPGA gas pedals, should be developed to further compress the evaluation latency to within 1 second and reduce the hardware cost. At the level of application expansion, it is proposed to build a multilingual cultural adaptation engine for the 'Belt and Road', which can support the DA of Chinese and Arabic language teaching.

Declarations

This work is supported by the Key Project of Humanities and Social Sciences Research in Higher Education Institutions of Anhui Provincial Department of Education (No. 2022AH052860 and No. 2024AH052914) and the Provincial Quality Project of Anhui Provincial Department of Education (No. 2022kcsz196).

The author declares that she has no conflicts of interest.

References

An, S., Zhang, S., Cai, Z. et al. (2024) 'Revealing the interplay of cognitive, meta-cognitive, and social processes in university students' collaborative problem solving: a three-stage analytical framework', *International Journal of Computer-Supported Collaborative Learning*, Vol. 13, pp.1–31.

Brasoveanu, A. and Dotlačil, J. (2021) 'Reinforcement learning for production-based cognitive models', *Topics in Cognitive Science*, Vol. 13, No. 3, pp.467–487.

Fujita, K., Nikaido, T., Burrow, M.F. et al. (2018) 'Effect of the demineralisation efficacy of MDP utilized on the bonding performance of MDP-based all-in-one adhesives', *Journal of Dentistry*, Vol. 77, pp.59–65.

Hosoda, C., Tanaka, K., Nariai, T. et al. (2013) 'Dynamic neural network reorganization associated with second language vocabulary acquisition: a multimodal imaging study', *Journal of Neuroscience*, Vol. 33, No. 34, pp.13663–13672.

Huang, Y., Peng, P., Zhao, Y. et al. (2023) 'Hierarchical adaptive value estimation for multi-modal visual reinforcement learning', *Advances in Neural Information Processing Systems*, Vol. 36, pp.46724–46736.

Ji, G.X., Chan, P.W.K., McCormick, A. et al. (2024) 'Scholarly responses to 'UNESCO Global Education Monitoring Report 2024 Pacific Technology in Education: a tool on whose terms?', *International Education Journal: Comparative Perspectives*, Vol. 23, No. 2, pp.154–170.

Jones, S.D. and Westermann, G. (2022) 'Under-resourced or overloaded? Rethinking working memory deficits in developmental language disorder', *Psychological Review*, Vol. 129, No. 6, p.1358.

Kim, N. and Nam, C.S. (2020) 'Adaptive control of thought-rational (ACT-R): applying a cognitive architecture to neuroergonomics', *Neuroergonomics: Principles and Practice*, Vol. 26, pp.105–114.

Kozulin, A. and Presseisen, B.Z. (1995) 'Mediated learning experience and psychological tools: Vygotsky's and Feuerstein's perspectives in a study of student learning', *Educational Psychologist*, Vol. 30, No. 2, pp.67–75.

Kuang, J., Shen, Y., Xie, J. et al. (2025) 'Natural language understanding and inference with MLLM in visual question answering: a survey', *ACM Computing Surveys*, Vol. 57, No. 8, pp.1–36.

Lorusso, M.L., Giorgetti, M., Travellini, S. et al. (2018) 'Giok the alien: an AR-based integrated system for the empowerment of problem-solving, pragmatic, and social skills in pre-school children', *Sensors*, Vol. 18, No. 7, p.2368.

Ma, C. (2025) 'China's Achievements in digital education in the wake of education Informatization 2.0 action plan', *Science Insights Education Frontiers*, Vol. 27, No. 1, pp.4435–4451.

Mattys, S.L., Davis, M.H., Bradlow, A.R. et al. (2012) 'Speech recognition in adverse conditions: a review', *Language and Cognitive Processes*, Vol. 27, Nos. 7–8, pp.953–978.

Mei, D. (2022) 'Understanding the nature and rationale of China's English curriculum – an explicit interpretation of English Curriculum Standards for Compulsory Education (2022 Edition)', *Journal of Teacher Education*, Vol. 9, No. 3, pp.104–111.

Mohammadi, M., Tajik, E., Martinez-Maldonado, R. et al. (2025) 'Artificial intelligence in multimodal learning analytics: a systematic literature review', *Computers and Education: Artificial Intelligence*, Vol. 6, p.100426.

Mu, S., Cui, M. and Huang, X. (2020) 'Multimodal data fusion in learning analytics: a systematic review', *Sensors*, Vol. 20, No. 23, p.6856.

Ochoa, X., Lang, A.C. and Siemens, G. (2017) 'Multimodal learning analytics', *The Handbook of Learning Analytics*, Vol. 1, pp.129–141.

Otto, A.R., Skatova, A., Madlon-Kay, S. et al. (2015) 'Cognitive control predicts use of model-based reinforcement learning', *Journal of Cognitive Neuroscience*, Vol. 27, No. 2, pp.319–333.

Peñarroja, M.R. (2021) 'Corpus pragmatics and multimodality: compiling an ad-hoc multimodal corpus for EFL pragmatics teaching', *International Journal of Instruction*, Vol. 14, No. 1, pp.927–946.

Posri, S. and Chansirisira, P. (2023) 'Components and indicators of participation in the implementation of quality assurance of small schools under the Office of the Basic Education Commission', *Journal of Education and Learning*, Vol. 12, No. 5, pp.74–84.

Qiu, S. (2024) 'Improving performance of smart education systems by integrating machine learning on edge devices and cloud in educational institutions', *Journal of Grid Computing*, Vol. 22, No. 1, p.41.

Qureshi, M.N.I., Oh, J., Cho, D. et al. (2017) 'Multimodal discrimination of schizophrenia using hybrid weighted feature concatenation of brain functional connectivity and anatomical features with an extreme learning machine', *Frontiers in Neuroinformatics*, Vol. 11, p.59.

Ritter, F.E., Tehranchi, F. and Oury, J.D. (2019) 'ACT-R: a cognitive architecture for modeling cognition', *Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 10, No. 3, p.e1488.

Shabani, K., Khatib, M. and Ebadi, S. (2010) 'Vygotsky's zone of proximal development: instructional implications and teachers' professional development', *English Language Teaching*, Vol. 3, No. 4, pp.237–248.

Wang, R., Xue, Y. and Wei, H. (2022) 'Cross-modal translation: a study of English subtitle translation from the perspective of multimodal interaction', *International Journal of Applied Linguistics and Translation*, Vol. 8, No. 1, pp.1–9.

Wang, S., Ni, L., Zhang, Z. et al. (2024) 'Multimodal prediction of student performance: a fusion of signed graph neural networks and large language models', *Pattern Recognition Letters*, Vol. 181, pp.1–8.

Zago, M., Luzzago, M., Marangoni, T. et al. (2020) '3D tracking of human motion using visual skeletonization and stereoscopic vision', *Frontiers in Bioengineering and Biotechnology*, Vol. 8, p.181.

Zellers, R., Lu, X., Hessel, J. et al. (2021) 'Merlot: multimodal neural script knowledge models', *Advances in Neural Information Processing Systems*, Vol. 34, pp.23634–23651.

Zhang, R-C., Lai, H-M., Cheng, P-W. et al. (2017) 'Longitudinal effect of a computer-based graduated prompting assessment on students' academic performance', *Computers & Education*, Vol. 110, pp.181–194.

Zou, R., Dechsubha, T. and Wang, Y. (2024) 'Mapping the semiotic landscape in education: language, multimodality, and educational transformation', *Language Related Research*, Vol. 15, No. 5, pp.283–311.