Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 16 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i16.62751

PAPER

A Multimodal Interaction Mechanism for Mobile English Learning in Higher Education

Ran Zhao (✉), Ning Dong

Hebei University of Chinese Medicine, Shijiazhuang, China

zhaoran118@163.com

ABSTRACT

Sensor-enabled mobile devices have provided new opportunities for interaction analysis and cognitive assessment in mobile English learning within higher education. However, existing approaches are constrained by insufficient data synchronization accuracy, limited flexibility in multimodal feature fusion, high computational costs associated with sequence modeling, and the lack of effective intervention mechanisms. To address these limitations, an edge-side multimodal interaction analysis and adaptive intervention framework was developed. Temporal alignment of heterogeneous sensor data was achieved, while lightweight cross-modal feature fusion was implemented using channel-attention-based representations. A bidirectional Mamba architecture was introduced for processing long-sequence data, and multitask learning was employed to simultaneously evaluate interaction quality and cognitive load. Behavioral evolution patterns were further explored through sequential pattern mining and cognitive network analysis. On this basis, a hierarchical feedback strategy was designed to enable intelligent intervention. The proposed framework provides both theoretical foundations and technical support for the development of intelligent mobile educational applications.

KEYWORDS

mining mobile learning, college English learning, multimodal interaction, feature fusion, sequence modeling, adaptive intervention, edge intelligence

1 INTRODUCTION

The widespread adoption of intelligent mobile devices has established mobile learning as an important platform for English language education in higher education institutions [1, 2]. Through the integration of multiple embedded sensors, continuous streams of learner behavioral data can be collected, providing the technological foundation for investigating interaction characteristics and cognitive dynamics throughout the learning process [3, 4]. Simultaneously, the advancement of mobile intelligent technologies has facilitated the evolution of educational applications toward edge-side intelligent systems in which perception, analysis, decision-making,

Zhao, R, Dong, N. (2026). A Multimodal Interaction Mechanism for Mobile English Learning in Higher Education. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(16), pp. 134–148. https://doi.org/10.3991/ijim.v20i16.62751

Article submitted 2026-05-13. Revision uploaded 2026-06-18. Final acceptance 2026-06-20.

and feedback are seamlessly integrated [5, 6]. Although mobile English learning applications have been extensively adopted, learner states cannot be comprehensively characterized through a single interaction modality alone [7, 8]. The present study possesses significance at theoretical, technological, and practical levels. From a theoretical perspective, the intrinsic relationships between multimodal behavioral patterns and cognitive states in mobile English learning environments can be systematically elucidated, thereby enriching the theoretical foundations of mobile learning analytics and multimodal human–computer interaction research [9, 10]. From a technological perspective, a lightweight framework for multimodal perception, modeling, and adaptive intervention can be developed under the computational and energy constraints of mobile devices, providing a transferable technical paradigm for the intelligent evolution of mobile educational applications [11, 12]. From a practical perspective, intelligent feedback mechanisms can be designed on the basis of discovered interaction patterns, thereby enhancing learner engagement, reducing cognitive load, and ultimately improving the quality of English learning in mobile environments [13, 14].

Despite the substantial progress achieved in mobile learning research, several critical limitations remain unresolved. At the data-processing level, most existing studies have relied on a single data source, such as audio signals or touch interactions, thereby failing to fully exploit the sensing capabilities provided by multimodal sensor environments [3, 10]. Temporal synchronization errors caused by inconsistent sampling frequencies across heterogeneous sensors have not yet been effectively addressed. Furthermore, edge-side solutions for data acquisition and preprocessing that can operate efficiently across mainstream mobile operating systems remain relatively scarce [5, 11]. At the feature-fusion level, conventional approaches have predominantly relied on feature concatenation or fixed-weight allocation strategies. Consequently, the relative contributions of different modalities cannot be adaptively adjusted in response to dynamic learning contexts, making it difficult to achieve an effective balance between fusion accuracy and computational overhead on mobile devices [9, 15]. At the levels of temporal modeling and state assessment, existing sequence modeling approaches exhibit limited suitability for real-time inference on mobile devices. When long sequential data are processed, gradient vanishing is frequently encountered in certain models, whereas excessive computational complexity in others imposes substantial processing burdens on resource-constrained devices [6, 12]. In addition, existing evaluation frameworks are predominantly limited to the classification of interaction states, while the quantitative assessment of cognitive load remains insufficiently addressed, resulting in a restricted evaluation scope and limited analytical depth [4, 16]. From the perspectives of behavioral mining and practical implementation, most studies have focused primarily on basic state recognition, whereas the temporal evolution patterns underlying interaction behaviors have not been thoroughly investigated. Furthermore, intervention approaches have largely been characterized by homogeneous designs, preventing the formulation of differentiated strategies based on learners' real-time interaction states. The practical effectiveness of existing intervention schemes has also rarely been validated through systematic experimental evaluation [8, 13].

To address the aforementioned limitations, a comprehensive framework for multimodal interaction analysis and adaptive intervention is developed. A cross-platform data acquisition and temporal synchronization scheme is designed to ensure compatibility across mainstream mobile operating systems. A lightweight cross-modal feature fusion module is constructed based on a channel-attention mechanism, enabling the dynamic and adaptive adjustment of modality contributions. To overcome the computational constraints associated with long-sequence modeling

on mobile devices, a bidirectional Mamba architecture is integrated with a multitask learning framework. Through this design, interaction quality assessment and cognitive load quantification are performed simultaneously. Building upon these capabilities, the temporal evolution patterns of interaction behaviors are further investigated across different learner groups, and a hierarchical intervention strategy is developed. The effectiveness and practical value of the proposed framework are validated through a series of experiments. The results demonstrate that robust deployment can be achieved across diverse commercial mobile learning platforms.

2 MULTIMODAL INTERACTION MECHANISM AND METHODOLOGY

To address the challenges associated with the accurate synchronization of heterogeneous multimodal data and the efficient long-sequence modeling in mobile learning environments, an edge-side framework for multimodal interaction analysis and cognitive state inference is developed. The overall processing pipeline of the proposed framework is illustrated in Figure 1. Four core modules are incorporated: (i) integrated edge-side data acquisition and temporal synchronization, (ii) lightweight feature fusion based on a channel-attention mechanism, (iii) multitask state inference using a bidirectional Mamba architecture, and (iv) temporal evolution analysis of interaction behaviors.

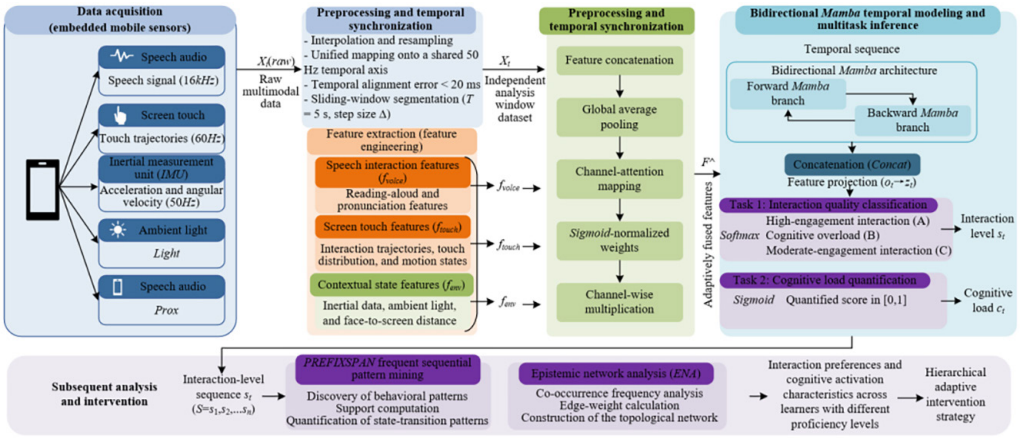

Fig. 1. Overall architecture of the edge-side framework for multimodal interaction analysis and state inference

2.1 Multimodal interaction data acquisition and edge-side temporal synchronization preprocessing

Accurate acquisition and temporal alignment of heterogeneous data constitute the foundation for interaction analysis and cognitive state inference. To accommodate the characteristics of mobile computing environments, an integrated edge-side sensing engine is developed for both Android and iOS platforms. By directly accessing embedded mobile sensors, comprehensive interaction data are collected without reliance on external sensing devices, thereby eliminating deployment constraints associated with additional hardware. Five categories of basic data are acquired simultaneously, including (1) speech audio, (2) screen-touch interactions, (3) nine-axis inertial measurement unit signals, (4) ambient light conditions, and (5) proximity

sensing data. For the *i*-th interaction event, the corresponding raw multimodal data can be represented as:

$$X_i^{(raw)} = (X_i^{(voice)}, X_i^{(touch)}, X_i^{(IMU)}, X_i^{(light)}, X_i^{(prox)}) \tag{1}$$

where, $X_i^{(raw)}$ denotes the complete data vector associated with the *i*-th interaction event. Each sub-vector represents the speech waveform, touch trajectory, inertial sensing signals, ambient light intensity, and proximity sensing status, respectively. To preserve the intrinsic characteristics of each modality, modality-specific sampling frequencies are adopted. Speech audio is sampled at 16 kHz to retain fine-grained spectral information relevant to speech production and pronunciation. Touch interaction data are collected at 60 Hz to capture continuous finger-movement trajectories with high temporal resolution. Inertial measurement unit signals are sampled at 50 Hz in accordance with commonly adopted operational specifications for mobile inertial sensors. Owing to hardware heterogeneity and device-specific configurations, the collected raw data are characterized by asynchronous temporal structures. If directly utilized for downstream modeling, such temporal discrepancies can substantially weaken cross-modal feature associations and influence the accuracy of subsequent state identification.

To eliminate temporal misalignment caused by inconsistent sampling frequencies among heterogeneous data sources, a hybrid calibration strategy combining cubic spline interpolation and linear resampling is employed. Cubic spline interpolation is first applied to fit discrete signals under the constraint of second-order continuous differentiability. For any two adjacent sampling nodes, $[x_m, x_{m+1}]$, the interpolation function $S(x)$ satisfies:

$$\begin{cases} S_m(x_m) = f(x_m), S_m(x_{m+1}) = f(x_{m+1}) \\ S'_m(x_{m+1}) = S'_{m+1}(x_{m+1}) \\ S''_m(x_{m+1}) = S''_{m+1}(x_{m+1}) \end{cases} \tag{2}$$

where, $f(x)$ denotes the original discrete signal generated by a sensor, x_m and x_{m+1} represent two adjacent physical sampling timestamps, and $S_m(x)$ denotes the piecewise interpolation function constructed within the corresponding interval. This approach can effectively preserve fine-grained signal characteristics while reducing interpolation-induced distortions. Following interpolation, linear temporal resampling is performed to project all modalities onto a shared 50 Hz temporal axis. As a result, the global temporal synchronization error is maintained below 20 ms, satisfying the temporal consistency requirements for behavior analysis in mobile learning environments. After data calibration, the continuous data stream is segmented through a sliding-window mechanism. Let T denote the window length and Δ denote the step size. The normalized dataset associated with the *t*-th independent analysis window can be expressed as:

$$X_t = \{X_{t,\tau}^{(voice)}, X_{t,\tau}^{(touch)}, X_{t,\tau}^{(IMU)}, X_{t,\tau}^{(light)}, X_{t,\tau}^{(prox)}\}_{\tau=1}^{T} \tag{3}$$

where, t denotes the analysis window, and τ denotes the sampling instant within the current window. The window length is fixed at 5 s, a configuration selected to accommodate the short-term attention distribution features typically observed during mobile English learning activities. This setting enables complete capture of individual interaction episodes while maintaining sufficient sensitivity to detect rapid fluctuations in cognitive states. Consequently, an effective balance is achieved between sample completeness and temporal sensitivity.

2.2 Multimodal feature extraction and lightweight fusion based on channel attention

The temporally synchronized and window-segmented raw sensor signals primarily reflect surface-level interaction behaviors. To characterize learners' interaction preferences and cognitive states, deeper behavioral information is extracted through feature engineering. Fine-grained feature extraction is conducted from three complementary dimensions: speech interaction, screen-touch interaction, and contextual state information. These dimensions collectively capture learners' spoken language production, interface operation patterns, and surrounding learning conditions. For the speech dimension, features associated with English reading-aloud and pronunciation activities are extracted, including Mel-frequency cepstral coefficients, fundamental frequency, speech rate, and pause-related indicators. Among these features, speech rate provides a direct measure of oral fluency and is computed as $r_{speech} = N_{syl}/T_{speech}$ where N_{syl} denotes the total number of syllables detected within a speech segment, and T_{speech} represents the effective speaking duration. All features are concatenated to form a fixed-dimensional speech feature vector. For the touch dimension, screen-interaction sequences are utilized to quantify operational trajectories, motion states, and click-distribution characteristics. To evaluate the spatial dispersion of touch interactions, information entropy is introduced and calculated as:

$$H_{click} = -\sum_{c=1}^{N_c} \frac{N_c^{(c)}}{N_{click}} \log_2 \frac{N_c^{(c)}}{N_{click}} \tag{4}$$

where, N_c denotes the total number of grid partitions defined on the screen interface, N_{click} represents the total number of click events within the current analysis window, and $N_c^{(c)}$ denotes the click frequency observed in the individual partition. Contextual state information is incorporated as complementary features. Device motion states are assessed using inertial measurement unit signals. The intensity of device perturbation is quantified through the signal magnitude area, defined as $\text{SMA} = \frac{1}{N}\sum_t |a_x(t)| + |a_y(t)| + |a_z(t)|$, where N denotes the total number of inertial samples within the analysis window, and $a_x(t)$, $a_y(t)$, and $a_z(t)$ represent the triaxial acceleration measurements at time t. Furthermore, the dominant frequency of the acceleration signal is estimated using the short-time Fourier transform. Ambient light intensity and user–device proximity information are also quantified. Finally, three modality-specific feature vectors are generated, denoted as f_{voice}, f_{touch}, and f_{env}, respectively. The dimensionality of each feature vector is determined by the number of features extracted from the corresponding modality.

A single modality can characterize learning states only from a limited perspective. To exploit the complementary information contained in heterogeneous modalities, an effective cross-modal feature fusion strategy is required. Conventional fusion approaches typically rely on direct feature concatenation or fixed-weight allocation schemes, which cannot dynamically adjust modality contributions according to changing learning tasks. Furthermore, the large number of parameters limits the applicability to resource-constrained mobile devices. To address these limitations, a channel-attention mechanism is introduced. A channel-compression strategy is further incorporated to reduce model complexity to reduce the computational cost of on-device inference while maintaining fusion accuracy. Initially, the three types of modality-specific features are concatenated to form a global joint feature representation $F = [f_{voice}, f_{touch}, f_{env}] \in R^{d_{all}}$, where d_{all} denotes the total number of channels after

multimodal feature concatenation. To aggregate channel-level feature information, global average pooling is applied:

$$z = GAP(F) = \frac{1}{d_{all}} \sum_{j=1}^{d_{all}} F_j \tag{5}$$

where, z denotes the global feature descriptor, and F_j represents the value of the j-th channel in the joint feature representation. The pooled feature vector is subsequently passed through a two-layer fully connected network to perform weight mapping. A compression-ratio constraint is introduced to reduce the dimensionality of the intermediate feature representation, thereby minimizing redundant parameters. The normalized attention weights are then obtained as $\alpha = \sigma(W_2 \cdot \delta(W_1 \cdot z))$, where W_1 and W_2 are learnable weight matrices, δ denotes the rectified linear unit activation function, and σ represents the Sigmoid function, which maps the output to the interval [0, 1] to achieve attention weight normalization. The resulting attention weights are employed to adaptively reweight the original joint features, enabling feature selection and enhancement $\tilde{F} = F \odot \alpha$, where $\odot$ denotes the Hadamard product along the channel dimension and $\tilde{F}$ represents the fused feature after adaptive modality-weight adjustment. This mechanism dynamically amplifies the contributions of highly informative modalities while suppressing redundant and less relevant information. Furthermore, the lightweight network architecture effectively supports real-time computation on resource-constrained mobile devices.

2.3 Bidirectional mamba-based multitask inference of interaction quality and cognitive states

The fused features form a continuous temporal sequence. In mobile learning environments, such sequences are typically characterized by long temporal spans. Conventional sequence networks exhibit notable limitations when deployed on mobile devices. Recurrent architectures are prone to gradient vanishing when processing long sequences, whereas the computational complexity of attention-based models increases quadratically with sequence length, making them unsuitable for low-latency execution on resource-constrained devices. To address these challenges, temporal feature modeling is performed using a state-space model. The continuous-time formulation of the model can be expressed as:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t) + Dx(t) \tag{6}$$

where, $x(t)$ denotes the input fused feature at time t; $h(t)$ represents the latent state vector; and $y(t)$ denotes the model output. Matrix A governs the dynamic evolution of the latent state, whereas matrices B, C, and D perform input projection, output projection, and direct input-to-output mapping, respectively. After discretization of the continuous-time system into a recursive state-update formulation, the overall computational complexity scales linearly with the sequence length. This property enables efficient deployment within the computational constraints of mobile devices. To fully exploit contextual dependencies within temporal data, a bidirectional encoding architecture is constructed. Let $\{\tilde{F}_1, \tilde{F}_2, \ldots, \tilde{F}_N\}$ denote the complete sequence of fused features. The forward branch processes the sequence in chronological order

to capture historical behavior information, whereas the backward branch traverses the sequence in reverse temporal order to model the influence of future states. The corresponding outputs can be expressed as:

$$o_t^{fwd} = Mamba_{fwd}(\bar{F}_1, \ldots, \bar{F}_t), \quad o_t^{bwd} = Mamba_{bwd}(\bar{F}_t, \ldots, \bar{F}_N) \tag{7}$$

The outputs of the two branches are subsequently integrated through feature concatenation, yielding temporal features enriched with contextual information $o_t = concat(o_t^{fwd}, o_t^{bwd})$.

Building upon the features (o_t) output by bidirectional encoding, a fully connected network is employed to perform nonlinear mapping of high-dimensional features. The mapping process can be expressed as:

$$z_t = W_3 \cdot ReLU(W_2 \cdot o_t + b_2) + b_3 \tag{8}$$

where, W_2 and W_3 denote learnable weight matrices, b_2 and b_3 represent bias vectors, and *ReLU* introduces nonlinear representational capacity into the network. The mapped feature representation is shared by two learning tasks. The first task involves interaction quality classification. The class probability distribution is generated through the Softmax function, denoted as $p_t = Softmax(z_t)$, and the classification error is evaluated using the cross-entropy loss L_{class}. The second task focuses on cognitive load quantification. Specifically, as for the linearly transformed features, a Sigmoid activation function is used to constrain the output to the interval [0, 1]. The result is $c_t^{\text{CogLoad}} = \sigma(w^T o_t + b)$. The regression accuracy is assessed using the mean squared error loss L_{reg}. A joint optimization strategy is adopted for the model. Accordingly, the overall loss function can be expressed as:

$$L_{total} = \lambda_1 L_{class} + \lambda_2 L_{reg} \tag{9}$$

The hyperparameters λ_1 and λ_2 are introduced to regulate the loss weights associated with the two tasks, thereby balancing the learning objectives of classification and regression. This enables the simultaneous inference of interaction states and cognitive load.

2.4 Multimodal interaction sequence analysis and cognitive network analysis

Following the quantification of interaction quality and cognitive load at the individual-window level, the analytical focus is extended from instantaneous state recognition to the investigation of long-term behavioral evolution. This extension enables a systematic examination of learners' state-transition characteristics throughout the mobile English learning process. A continuous behavioral sequence is constructed using window-level interaction quality labels as the fundamental analytical units, denoted as $S = (s_1, s_2, \ldots, s_N)$, where N denotes the total number of windows observed during a complete learning session, and $s_t \in \{A, B, C\}$ represents the interaction level at the current time step. The three categories correspond to high-engagement interaction, moderate-engagement interaction, and cognitive overload, respectively. During data preprocessing, prolonged inactive periods characterized by the absence of meaningful learner interactions are removed to minimize the influence of non-informative samples on subsequent analyses. The PrefixSpan algorithm is employed to mine high-frequency behavior subsequences.

Pattern selection is performed based on the support measure, and the equation is defined as:

$$supp(p) = \frac{count(p)}{Total(S)} \tag{10}$$

where, $count(p)$ denotes the number of learner samples containing the target subsequence p; and $Total(S)$ represents the total number of behavioral sequences within the dataset. A minimum support threshold of $min\_supp = 0.05$ is adopted to identify representative behavioral patterns with sufficient generalizability. Interaction-state transition dynamics are quantified across different learner groups, thereby revealing characteristic behavioral evolution trajectories and interaction preferences.

Although frequent sequential pattern mining can reveal temporal ordering relationships among interaction states, it cannot quantify the strength of co-occurrence relationships or the underlying coordination mechanisms among different states. To address this limitation, epistemic network analysis is incorporated to transform discrete interaction states into a topological network representation, thereby providing a more comprehensive characterization of learners' latent cognitive activation patterns. The three interaction quality levels are defined as fixed network nodes. Co-occurrence frequencies among nodes are estimated using a sliding co-occurrence window. A temporal decay coefficient is introduced to reduce the influence of distant observations. The edge weight is computed as:

$$w_{ij} = \sum_{t}^{T} exp(-\eta \cdot \Delta t) \cdot I(s_t = i, s_{t+1} = j) \tag{11}$$

where, w_{ij} denotes the connection weight between nodes i and j; η represents the temporal decay coefficient; and Δt denotes the temporal interval between samples. I denotes a binary indicator function used to determine whether two states are simultaneously activated at adjacent time steps. Following the construction of the network matrix, principal component analysis is applied to reduce the high-dimensional network parameters into a two-dimensional feature space. Subsequently, the network centroids of different participant groups are computed, and statistical hypothesis testing is conducted. This approach enables the identification of distinctive characteristics among learners with different levels of English proficiency in terms of cognitive coupling structures and state-activation patterns. Consequently, the underlying mechanisms governing multimodal interactions in mobile English learning environments are comprehensively elucidated.

3 EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

3.1 Experimental setup

At the hardware level, representative mobile devices covering the mainstream consumer market were included. Both Android and iOS platforms were considered, and the devices were categorized into three performance tiers: entry-level, mid-range, and flagship models. This configuration was designed to reflect diverse real-world usage scenarios encountered by users. The entry-level category included Android devices equipped with a Qualcomm Snapdragon 695 processor and 8 GB of memory. The mid-range category consisted of Android devices powered by a Qualcomm Snapdragon 8+ processor with 12 GB of memory, together with the iPhone 13. The flagship category included Android devices equipped with a

Qualcomm Snapdragon 8 Gen 3 processor and 16 GB of memory, as well as the iPhone 15 Pro. At the software level, a cross-platform edge-side data acquisition engine was developed using Android Studio and Xcode to support multimodal sensor data collection and preprocessing. The deep learning framework was deployed on mobile devices through PyTorch Mobile with model quantization, enabling efficient inference under resource-constrained mobile computing environments.

A total of 186 undergraduate students with varying levels of English proficiency were recruited as participants. Based on their performance in the institutional English proficiency examination, participants were categorized into low-, intermediate-, and high-proficiency groups. All participants completed three representative mobile English learning activities within the dedicated experimental application: listening-and-repeat tasks, reading-comprehension tasks, and vocabulary-memorization tasks. A total of 21,740 valid analysis windows were obtained. The dataset comprised five categories of multimodal data, including speech audio, touch interaction data, nine-axis inertial measurement unit signals, ambient light information, and proximity sensing data. Labels were generated using the semi-supervised framework. Specifically, learning performance improvement and cognitive workload measures obtained from the NASA Task Load Index (NASA-TLX) were jointly incorporated into a clustering-based soft-label generation process. To reduce potential bias introduced by a single labeling source, random manual inspections were performed.

3.2 Comparative evaluation of temporal modeling architectures

Conventional temporal modeling architectures exhibit notable limitations when deployed for real-time inference on long interaction sequences in mobile learning environments. To evaluate the effectiveness of the proposed bidirectional Mamba architecture, a comparative study was conducted against several representative sequence-modeling approaches. The objective was to assess both predictive performance and deployment efficiency. All models were configured with identical input feature dimensions, training epochs, and learning rates. Five temporal architectures were implemented, including long short-term memory, gated recurrent unit, unidirectional Transformer, bidirectional Transformer, and the proposed bidirectional Mamba architecture. After training, all models were quantized and deployed on a mid-range Android device. Both predictive performance metrics and mobile resource-consumption indicators were recorded.

Table 1. Comparison of predictive performance and mobile resource consumption across temporal models

Model	Parameters (K)	Accuracy (%)	F1-Score (%)	Mean Absolute Error	Inference Latency (ms)	Peak Central Processing Unit Utilization (%)
Long short-term memory	142.6	81.36	80.92	0.127	18.63	22.15
Gated recurrent unit	116.8	82.15	81.75	0.121	16.27	20.86
Unidirectional Transformer	285.3	84.62	84.18	0.096	35.14	38.72
Bidirectional Transformer	567.5	85.79	85.26	0.089	62.58	45.36
Proposed bidirectional Mamba architecture	103.2	86.24	85.91	0.085	11.35	16.42

As shown in Table 1, the proposed bidirectional Mamba architecture required only 103.2 K parameters, representing the lightweight advantage among all evaluated approaches. In terms of predictive performance, a classification accuracy of 86.24% and an F1-score of 85.91% were obtained, both representing the highest values among the competing models. For the cognitive load regression, the mean absolute error was reduced to 0.085, outperforming even the bidirectional Transformer architecture. These results indicate that the bidirectional Mamba architecture is more effective in capturing long-sequence interaction features and contextual relationships. In terms of mobile deployment efficiency, the average inference latency per window was reduced to 11.35 ms, representing a reduction of 51.23 ms relative to the bidirectional Transformer and corresponding to a latency decrease of more than 80%. Peak central processing unit utilization was limited to 16.42%, which was consistently lower than that observed for both recurrent neural network and attention models. These findings can be attributed to the linear computational complexity of the Mamba architecture. By avoiding the quadratic complexity associated with Transformer-based architectures while simultaneously alleviating the long-sequence modeling limitations of traditional recurrent architectures. Consequently, the proposed architecture satisfies the real-time inference requirements of mobile environments.

3.3 Ablation study of the cross-modal fusion module

An ablation study was conducted to evaluate the effectiveness of the proposed channel-attention-based fusion module and to quantify the relative contributions of the speech, touch-interaction, and contextual-state modalities. The objective was to determine the influence of individual components on model performance. A total of six experimental configurations were evaluated, including two conventional fusion strategies and four modality-removal variants. The complete channel-attention-based fusion framework was used as the experimental model. All configurations employed the same temporal encoding network and multitask learning framework, thereby eliminating potential confounding effects introduced by other variables.

As shown in Table 2, the conventional feature concatenation and fixed-weight fusion schemes exhibited comparatively limited performance, with classification accuracies achieving 80.15% and 81.67%. In addition, relatively large mean absolute errors were observed. Using the proposed channel-attention-based fusion mechanism, the accuracy increased to 86.24%, and the mean absolute error was reduced to 0.085. These results demonstrate that dynamic weighting effectively enhances informative features while suppressing redundant or less relevant information, thereby improving the overall effectiveness of fusion. In terms of the relative importance of individual modalities, the performance degradation (79.58%) occurred when the touch-interaction modality was excluded. Under this condition, the mean absolute error increased to 0.142, indicating the largest performance degradation. These findings indicate that touch interaction data serve as the primary source of information for inferring learner states. When either the speech or contextual-state modality was removed, the model performance also declined to varying degrees, with classification accuracies decreasing to 82.31% and 83.74%, respectively. This result confirms that all three modalities contribute unique and complementary information that cannot be fully substituted by the others. In particular, contextual-state data provide additional information regarding device operating conditions and external contextual factors, thereby enriching the representation of learner states.

Table 2. Results of the ablation study on the cross-modal fusion module

Group	Configuration	Accuracy (%)	F1-Score (%)	Mean Absolute Error
Baseline 1	Direct feature concatenation	80.15	79.63	0.135
Baseline 2	Fixed-weight fusion	81.67	81.24	0.124
Baseline 3	Without the speech modality	82.31	81.85	0.118
Baseline 4	Without the touch-interaction modality	79.58	79.11	0.142
Baseline 5	Without the contextual-state modality	83.74	83.26	0.107
Proposed model	Channel-attention-based adaptive fusion	86.24	85.91	0.085

3.4 Comparative evaluation of single-modality and multimodal sensing strategies

A comparative experiment was conducted to evaluate the effectiveness of single-modality and multimodal sensing strategies. The objective was to assess the practical value of the proposed multi-sensor data acquisition framework and to determine the suitability of multimodal interaction analysis for mobile English learning environments. Seven experimental configurations were established, covering single-modality sensing, bimodal sensing combinations, and full multimodal fusion. Identical model architectures and training parameters were employed across all experimental groups. The representational capability of each modality combination was evaluated using both classification and regression metrics.

The results presented in Figure 2 clearly demonstrate the performance gains achieved through multimodal fusion. Among the three single-modality configurations, the touch-interaction modality exhibited the strongest representational capability, achieving a classification accuracy of 74.61%. In contrast, the contextual-state modality yielded the weakest performance, with an accuracy of only 68.92%. Furthermore, the mean absolute error exceeded 0.17 for all single-modality configurations, indicating substantial limitations in the ability of individual modalities to characterize learners' states comprehensively. Superior performance was consistently observed for all bimodal configurations relative to their single-modality counterparts. Among the bimodal combinations, the integration of speech and touch-interaction modalities produced the best results, achieving an accuracy of 82.57%. This finding suggests that strong complementary relationships exist between spoken language production and interface interaction behaviors during mobile English learning activities. The highest performance was obtained when all three modalities were jointly incorporated. Under the full multimodal fusion configuration, compared with the best-performing single-modality configuration, the accuracy increased by 11.63 percentage points, while the mean absolute error decreased from 0.172 to 0.085. These improvements indicate that substantial complementary information is provided by the different sensing modalities. Speech, touch interaction, and contextual-state information capture distinct aspects of learner behavior and learning context. The heterogeneous information sources contribute to the construction of a more comprehensive feature representation space.

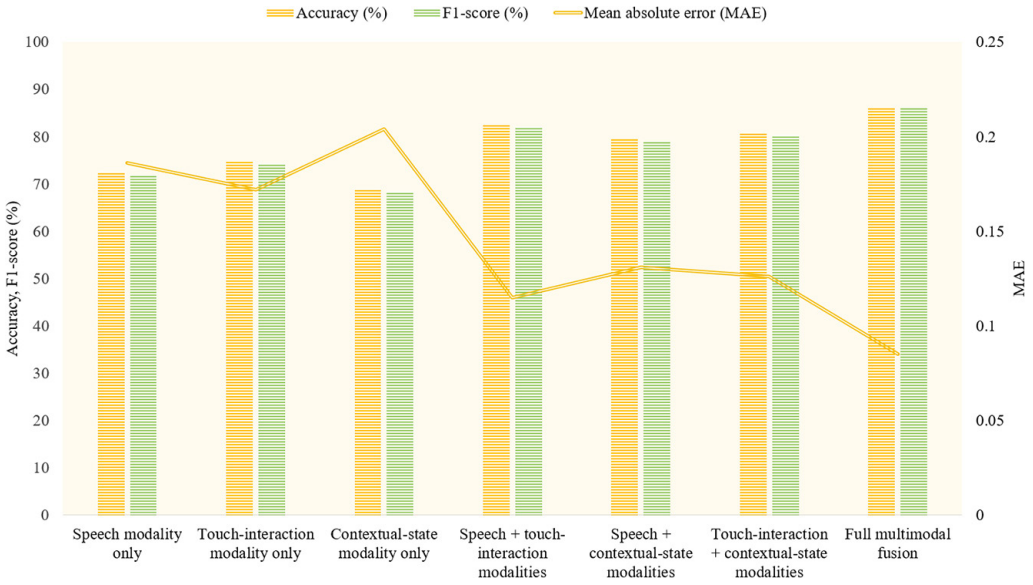

Fig. 2. Comparison of model performance across different modality configurations

3.5 End-to-end mobile deployment performance evaluation

From an engineering deployment perspective, the stability and resource consumption of the proposed data processing, feature fusion, and state inference framework were evaluated during prolonged operation on mobile devices with different hardware specifications. The objective was to assess the feasibility of large-scale commercial deployment. To this end, entry-level, mid-range, and flagship Android and iOS devices were selected for experimentation. The complete algorithmic pipeline was executed continuously for one hour on each device. During the evaluation, the average inference latency, hardware resource utilization, sensor synchronization error, and device power consumption were continuously monitored and recorded. The mean values of these performance indicators were subsequently calculated for comparative analysis.

As shown in Table 3, stable operation was achieved across mobile devices of different performance tiers and operating systems. The maximum temporal synchronization error observed during testing was 17.36 ms, which remained below the predefined threshold of 20 ms. Consequently, adequate temporal consistency was maintained for multimodal data. In terms of inference efficiency, an average latency of 15.62 ms was observed on the entry-level Android device, indicating that real-time interaction analysis remained feasible. The lowest latency was achieved on the flagship device, where the average inference time was reduced to 7.21 ms, demonstrating excellent responsiveness for edge-side deployment. In terms of resource consumption, peak memory usage of the complete pipeline algorithm did not exceed 186.54 MB, remaining well within the typical memory allocation limits of contemporary mobile applications. Similarly, average central processing unit utilization remained below 22% across all evaluated devices, indicating that the execution is unlikely to interfere with foreground user activities. A gradual reduction in power consumption was observed as device performance increased. The highest average power consumption was recorded on the entry-level Android device at 1.26 W. The value suggests that prolonged operation can be sustained without introducing excessive thermal load or abnormal battery depletion. Only minor

differences were observed between Android and iOS devices of comparable performance tiers. This finding demonstrates strong cross-platform compatibility and satisfies the practical deployment requirements of commercial mobile learning applications.

Table 3. Runtime performance of the complete algorithmic pipeline across mobile devices

Device Tier	Operating System	Average Latency (Ms)	Average Central Processing Unit Utilization (%)	Memory Usage (Mb)	Temporal Synchronization Error (Ms)	Average Power Consumption (W)
Entry-level	Android	15.62	21.35	186.54	17.36	1.26
Mid-range	Android	11.35	16.42	152.37	16.82	1.03
Flagship	Android	7.58	12.17	121.65	15.95	0.87
Mid-range	iOS	10.86	15.73	146.28	16.51	0.98
Flagship	iOS	7.21	11.64	118.32	15.67	0.82

4 CONCLUSIONS AND FUTURE DIRECTIONS

A comprehensive edge-side framework integrating multimodal sensing, cognitive state inference, and adaptive intervention was developed for mobile English learning in higher education. The experimental results demonstrate that the proposed channel-attention-based lightweight fusion module dynamically adjusts modality contributions according to learning context, thereby strengthening feature coupling among heterogeneous data sources while maintaining low computational overhead for edge-side deployment. Furthermore, the bidirectional Mamba-based multitask learning framework effectively addresses the limitations of conventional temporal models in long-sequence modeling by leveraging its linear computational complexity. Specifically, it mitigates the challenges of insufficient modeling accuracy and excessive computational overhead commonly encountered in traditional sequence-learning architectures, thereby enabling real-time inference on Android and iOS devices across a wide range of hardware specifications. By integrating sequential pattern mining with cognitive network analysis, the present study elucidates the intrinsic relationships between learner interaction behaviors and cognitive load and identifies several representative state-transition patterns. Application-level comparative experiments further demonstrate that the proposed hierarchical intervention strategy can accurately align instructional support with the needs of learners at different interaction levels. As a result, the proportion of high-quality interactions is significantly increased, unnecessary cognitive expenditure is reduced, and the overall effectiveness of mobile English learning is substantially enhanced.

5 REFERENCES

[1] M. AbdAlgane, M. A. Elkot, and R. Ali, “Embracing mobile-based spaced learning to enhance English scientific vocabulary mastery and academic performance among ESP students,” *International Journal of Interactive Mobile Technologies*, vol. 19, no. 12, pp. 103–120, 2025. https://doi.org/10.3991/ijim.v19i12.54815

[2] O. Synekop, I. Lytovchenko, Y. Lavrysh, and V. Lukianenko, "Use of Chat GPT in English for engineering classes: Are students' and teachers' views on its opportunities and challenges similar?" *International Journal of Interactive Mobile Technologies*, vol. 18, no. 3, pp. 129–146, 2024. https://doi.org/10.3991/ijim.v18i03.45025

[3] S. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review," *Sensors*, vol. 20, no. 23, p. 6856, 2020. https://doi.org/10.3390/s20236856

[4] C. Larmuseau, J. Cornelis, L. Lancieri, P. Desmet, and F. Depaepe, "Multimodal learning analytics to investigate cognitive load during online problem solving," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1548–1562, 2020. https://doi.org/10.1111/bjet.12958

[5] F. Wang, M. Zhang, X. Wang, X. Ma, and J. Liu, "Deep learning for edge computing applications: A state-of-the-art survey," *IEEE Access*, vol. 8, pp. 58322–58336, 2020. https://doi.org/10.1109/ACCESS.2020.2982411

[6] D. L. Xu, T. Li, Y. Li, X. Su, S. Tarkoma, and T. Jiang, "Edge intelligence: Empowering intelligence to the edge of network," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1778–1837, 2021. https://doi.org/10.1109/JPROC.2021.3119950

[7] V. N. Hoi and G. M. Mu, "Perceived teacher support and students' acceptance of mobile-assisted language learning: Evidence from Vietnamese higher education context," *British Journal of Educational Technology*, vol. 52, no. 2, pp. 879–898, 2020. https://doi.org/10.1111/bjet.13044

[8] Y. Luo and M. Watts, "Exploration of university students' lived experiences of using smartphones for English language learning," *Computer Assisted Language Learning*, vol. 37, no. 4, pp. 608–633, 2022. https://doi.org/10.1080/09588221.2022.2052904

[9] M. Cukurova, M. Giannakos, and R. Martinez-Maldonado, "The promise and challenges of multimodal learning analytics," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1441–1449, 2020. https://doi.org/10.1111/bjet.13015

[10] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, "A review on data fusion in multimodal learning analytics and educational data mining," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 4, 2022. https://doi.org/10.1002/widm.1458

[11] Q. Chen, Z. Wang, Y. Su, L. Fu, and Y. Wei, "Educational 5G edge computing: Framework and experimental study," *Electronics*, vol. 11, no. 17, p. 2727, 2022. https://doi.org/10.3390/electronics11172727

[12] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proceedings of the IEEE*, vol. 107, no. 8, pp. 1738–1762, 2019. https://doi.org/10.1109/JPROC.2019.2918951

[13] Y. Dai and Z. Wu, "Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study," *Computer Assisted Language Learning*, vol. 36, nos. 5–6, pp. 861–884, 2021. https://doi.org/10.1080/09588221.2021.1952272

[14] O. Viberg, A. Kukulska-Hulme, and W. Peeters, "Affective support for self-regulation in mobile-assisted language learning," *International Journal of Mobile and Blended Learning*, vol. 15, no. 2, pp. 1–15, 2023. https://doi.org/10.4018/IJMBL.318226

[15] U. Bin Qushem, A. Christopoulos, S. Oyelere, H. Ogata, and M. Laakso, "Multimodal technologies in precision education: Providing new opportunities or adding more challenges?" *Education Sciences*, vol. 11, no. 7, p. 338, 2021. https://doi.org/10.3390/educsci11070338

[16] X. Yang and J. Hu, "Distinctions between mobile-assisted and paper-based EFL reading comprehension performance: Reading cognitive load as a mediator," *Computer Assisted Language Learning*, vol. 37, no. 7, pp. 2051–2082, 2022. https://doi.org/10.1080/09588221.2022.2143527

6 AUTHORS

Ran Zhao is a Lecturer at the Humanities and Management Department of Hebei University of Chinese Medicine, China. Her research interests include college English teaching, foreign linguistics, and applied linguistics. As the project leader, she has undertaken several scientific research projects at the provincial department level and school level, with more than six papers published and one book published (E-mail: zhaoran118@163.com).

Ning Dong is a master working on English teaching and English translation in the School of Humanities and Management at Hebei University of Chinese Medicine, China. Her research interests include English translation, English teaching, Cultural contrast, more than 5 papers published, and 3 books published (E-mail: dongning000521@126.com).