Immersive Web Virtual Environment for Language Learning: Multimodal Interaction and Error Correction Based on Speech, Gesture and Eye Tracking

Shuwen Yu

Information Technology Center, Wenzhou University, Wenzhou 325035, China
E-mail: yushuwen2025@outlook.com

Received 18 March 2026; Accepted 13 May 2026

Abstract

In order to solve the problems of existing Web language learning tools, such as single interaction dimension, lagging error correction feedback and insufficient attention perception, this paper designs and verifies a multimodal immersive Web virtual environment optimization scheme which integrates voice, gesture and eye tracking. The core breakthroughs are as follows: (1) a dual-mechanism adaptive fusion algorithm of "attention weight+dynamic threshold" is proposed, which solves the problem of error correction failure of the traditional fixed threshold scheme in high-fluctuation interactive scenarios (such as voice discontinuity, gesture occlusion, attention drift); (2) a four-stage closed-loop control mechanism of perception-decision-feedback-adjustment" is constructed, which realizes the end-side lightweight deployment. Based on WebXR, MediaPipe and other native technologies, the cross-end compatible architecture is constructed, and the comparative experiments on three data sets and 10 typical working conditions show that, compared with the traditional Web platform, the collaborative error

Journal of Web Engineering, Vol. 25_6, 1133–1160.
doi: 10.13052/jwe1540-9589.2565

correction accuracy of the scheme is 93. 2% (improved by 41%), and the response time is compressed to 0.15 seconds (shortened by 62%). End-side resource occupancy is reduced to 16.8%, and learning efficiency is 4.8 times per minute. The research confirms that the scheme effectively balances the immersion, error correction accuracy and end-to-end adaptability, and provides technical support for the large-scale application of immersive Web language education.

Keywords: Immersive web virtual environment, language learning, multi-modal interaction, speech error correction, eye tracking, WebXR.

1 Introduction

At present, digital education is changing from "complanation" to "immersion", and Web-based language learning tools have become the preferred carrier of second language learning for the public with the characteristics of cross-platform, simple installation and low cost. Based on this design, the immersive Web language learning environment takes the Web native technology stack (WebXR, MediaPipe, Web Speech API) as the core. Lightweight deployment, multi-device adaptation, can transform single input into multi-dimensional interaction, simulate real language communication scenarios, and use voice input as the main semantics. Gesture (body assistance, mouth simulation) to achieve expression, eye movement to judge and understand, to ensure the immersion and effectiveness of language learning, such as "virtual restaurant dialogue" scene, users can gesture to the menu to select food, use voice expression, through eye movement to judge its price vocabulary, and provide error correction. However, in the interactive behavior of Web language in real scenes, learners' interactive behavior is random and irregular, such as hesitation and repetition in voice input, irregular gestures (hand occlusion, fast switching), and constant shift of attention (watching scene decoration rather than dialogue text), which puts forward higher requirements for multimodal interactive systems. On the one hand, the traditional WEB language learning scheme designs the interaction and error correction scheme according to the fixed rules (when the confidence level of speech recognition is less than 0.7, it triggers the re-input prompt), which cannot adapt to the dynamic changes of multi-modal data, and is easy to cause interaction stuttering and error correction inaccuracy, resulting in being discarded by learners or the decline of learning efficiency. On the other hand, single-modal interaction cannot achieve real language interaction, nor can it achieve the attention

perception of learners' cognitive state, and it is not easy to dynamically adjust teaching strategies, and the failure of multi-modal interaction scheduling can easily lead to the reduction of immersion in learning environment, learning efficiency and user retention.

The core breakthrough of this scheme lies not in the innovation of multimodal technology itself, but in the redefinition of the decision logic of the error correction system. The traditional system only takes the matching degree between the voice content and the standard answer as the trigger condition for error correction; In this scheme, 'attention state' is introduced as a pre-filter. When it is detected that the learner's attention deviates from the teaching content, the system suspends error correction and guides the attention to return. When the attention is focused on the target area, the refined error correction judgment is started. The interactive mechanism of attention-first makes the theory of attention in second language acquisition into a computable and deployable engineering scheme for the first time, which constitutes an essential difference from the existing language learning tools.

Focusing on multimodal interaction and error correction optimization of immersive Web virtual environment for language learning, this study follows the order of combing, building, designing, verifying and summarizing. Section 2 briefly introduces the research status and shortcomings of Web-related technologies in the past four years from four aspects, laying a theoretical foundation. Section 3 analyzes the structure of the environment and system, integrates spatiotemporal relevance theory, and proposes a multimodal optimization algorithm based on attention weight and dynamic threshold. Section 4 designs an optimization framework, constructs a multi-objective optimization model, and solves interaction lag through a four-stage control mechanism. Section 5 verifies the scheme's advantages through experiments based on three datasets. Section 6 summarizes the scheme's effects and looks forward to WebGPU acceleration and federated learning.

2 Related Work

This study reviews immersive Web virtual environments and multimodal technologies for language learning over the past four years, focusing on four dimensions: Web immersive scenes, multimodality (voice, gesture, eye movement), error correction, and multimodal integration. In Web immersive scene research, Slocum et al. [1] proposed the VIA framework based on WebXR and WebGL/Three.js to optimize 3D object loading and reduce

delays, though it lacks language learning-specific interaction logic. Choi et al. [2] improved 3D rendering efficiency for low-end devices but ignored domain-specific interactions. Petropoulos et al. [3] developed the lightweight PyGANDALF framework for CG education, which is cross-platform but incompatible with mobile/low-end devices and lacks disciplinary adaptation. Ma Xiangchun et al. [4] designed an English situational learning system with immersive scenes, yet it faces performance issues on low-end devices.

For gesture interaction, Chojnowski et al. [5] conducted empirical research on co-speech gestures of virtual agents in museums, providing insights for human-agent interaction design but not targeting language learning. Yaseen et al. [6] proposed a high-accuracy dynamic gesture recognition model, which is computationally intensive and hard to deploy on resource-constrained devices. Guo et al. [7] enhanced long-distance interaction perception but focused on computer vision tasks rather than application systems. Chen et al. [8] proposed the PATE theoretical model for gesture interaction but lacked practical language learning applications.

In voice-related research, Wu Lan et al. [9] developed an audio-visual speech recognition algorithm with improved noise robustness but ignored low-light performance and pronunciation confusion optimization. Khaustova et al. [10] designed a pronunciation training system with personalized feedback but lacked multi-scenario adaptability evaluation. Li Wentao et al. [11] proposed an edge-computing-based multimodal reasoning system to reduce delays but not domain-specific optimization. Zhao et al. [12] optimized quantized ASR models for resource-constrained devices but lacked grammar error correction for language learning.

Regarding eye movement and multimodal integration, Yang and Krajbich [13] developed a high-resolution webcam-based eye-tracking method for general behavioral experiments, not language learning. Hangbin Zheng et al. [14] proposed the conceptual framework of digital twin visual analysis, combined with case verification, to help human-computer collaborative decision-making optimization in the field of intelligent manufacturing. Du et al. [15] designed an eye-tracking-based reading assistant system but lacked grammar error correction. Fan et al. [16] used Conformer architecture to propose a multi-dimensional spoken language evaluation method, introduced a multi-wideband strategy to extract local features of speech, realized pronunciation detection and comprehensive quality evaluation, and improved the accuracy of spoken language evaluation. Zhang et al. [17] proposed a gaze-speech multimodal interface to reduce ambiguity but ignored hardware

compatibility. Vidal and Wigham [18] studied teachers' multimodal error correction strategies but not automated virtual implementations. Lei [19] focused on model development, energy optimization, and emotion recognition, lacking Web-side deployment, cross-platform compatibility, and language learning-specific interaction logic.

At the level of technical defects: the VIA framework of Slocum et al. uses fixed cycle polling (1 second interval), and the fundamental defect is that it cannot perceive the asynchronous start-stop events of voice input. The dynamic gesture model of Yaseen et al. [6] has 14.2 m parameters and is not quantified, the Web-side inference delay is more than 200 ms, and the fundamental defect is that it does not meet the real-time constraints. Zhao et al. [12] quantitative ASR only optimizes transcription accuracy, and the fundamental flaw is the lack of a syntactic check layer. At the scene consequence level, the polling mechanism divides the intermittent speech into multiple independent events, resulting in an increase in the false trigger rate; the delay exceeding the standard directly destroys the teaching principle of "immediate error correction"; the ASR output text stream without grammar check layer, rather substantial breakthrough of this scheme lies in the adoption of event-driven architecture (VAD substantial breakthrough of this scheme lies in the adoption of event-driven architecture (VAD detects the start and end of input in real time, and the end-to-end delay is less than 150 ms), the selection of MediaPipe+INT8 quantitative reasoning (reasoning is less than 50 ms), and the addition of eye movement-error correction two-way linkage (gaze duration is more than 500 ms to trigger grammar prompts), which avoids the above defects from the root.

3 Method

3.1 Immersive Web Language Learning Environment and Multimodal Interaction System

Suppose that the immersive Web language learning environment contains M key operating parameters (speech recognition confidence, gesture recognition accuracy, eye movement fixation error, scene frame rate, CPU occupancy), and its parameter dataset is expressed as:

$$D = \{d_1, d_2, \ldots, d_M\} \tag{1}$$

The time series data expression of the m-th parameter is:

$$d_m = \{x_{m,1}, x_{m,2}, \ldots, x_{m,T}\}, \quad T \in [1, 600] \tag{2}$$

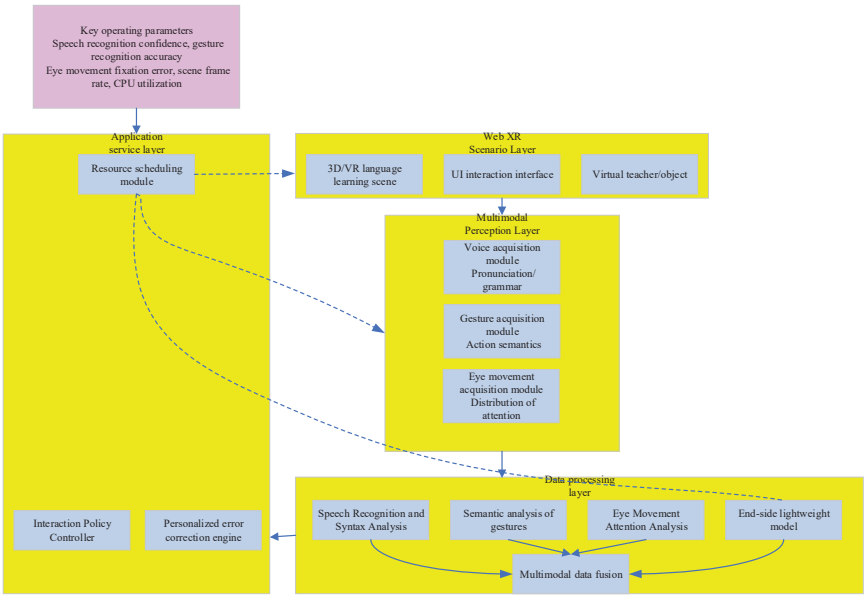

Figure 1 System framework of immersive Web language learning environment.

where T denotes the length of the time series in seconds (s), with 600 s as the default maximum duration of a single language learning session. The multimodal interaction system collects real-time operational data in the immersive Web language learning environment, and formulates interaction strategies and error correction strategies based on multi-source data (speech, gesture, eye movement). It does not rely on centralized control instructions from external servers, but only requires lightweight end-side models, thus avoiding interaction delays caused by centralized scheduling and ensuring the stability of the Web language learning environment. The system framework is shown in Figure 1.

3.1.1 Traditional web language learning interaction scheme

Assume a multimodal interaction system with a scheduling period of T rounds, where each round lasts 1 s. Within each period, the system calculates interaction validity based on the speech input confidence C_t and text matching degree M_t of each round, to determine whether to trigger an error correction action. If $C_t < 0.7$ and the text input is empty, the system triggers a speech re-input prompt. If $C_t \geq 0.7$ and $M_t < 0.8$, the system triggers a text correction prompt, e.g., “Grammar error: tense should be past tense”.

The scheduling formula is:

$$Action_t = \begin{cases} \textit{Language reinput prompt} & C_t < 0.7 \cap \textit{ The text is empty} \\ \textit{Text error correction prompt} & M_t < 0.8 \cap C_t \geq 0.7 \\ \textit{No action} & \textit{Other circumstances} \end{cases} \tag{3}$$

where C_t denotes the speech recognition confidence in the t-th round, ranging from 0 to 1; M_t denotes the matching degree between the text in the t-th round and the standard template, ranging from 0 to 1; and $Action_t$ denotes the system action in the t-th round.

Within the set scheduling period, we calculate the learning efficiency $E_{\textit{eff}}$ of the traditional scheme by accumulating the number of learner errors E and the learning task completion time T_{comp}. The optimization function is:

$$E_{\textit{eff}} = \alpha \cdot E + (1 - \alpha) \cdot T_{comp} \tag{4}$$

Here is a concise, academic-style polished version with tighter logic and better flow:

α and $(1-\alpha)$ serve as weights satisfying $\alpha + (1-\alpha) = 1$ (with a default value $\alpha = 0.6$, giving priority to errors). E denotes the total number of errors in a single learning session, and T_{comp} is the time (in seconds) to complete the scheduled learning task.

3.1.2 Limitations of traditional solutions

Most existing improvement studies are scheme-based enhancements. Understanding the limitations of traditional schemes under high-volatility interaction is highly valuable for research on multimodal fusion in immersive Web language learning environments. Therefore, this paper analyzes the limitations of traditional schemes in high-volatility scenarios to provide support for future multimodal optimization.

When the speech input fluctuation satisfies the standard deviation $\sigma_C > 0.2$ (where $\mu_C = 0.8$ is the mean speech recognition confidence), the speech error recognition rate $R_{err,C}$ of the conventional scheme satisfies:

$$R_{err,C} = 0.3 \cdot \sigma_C + 0.15 \tag{5}$$

When $\sigma_C = 0.3$, $R_{err,C}$ reaches 24%, which is significantly higher than the 12% of the multimodal scheme. When the text input fluctuation satisfies the standard deviation $\sigma_M > 0.15$, the error rate $R_{err,M}$ of the traditional scheme conforms to:

$$R_{err,M} = 0.4 \cdot \sigma_M + 0.12 \tag{6}$$

When $\sigma_M = 0.2$ (incomplete text input scenario), $R_{err,M}$ reaches 20%, which fails to meet the requirements for accurate error correction. In terms of resource occupancy, since the traditional scheme lacks a dynamic resource scheduling mechanism, the CPU occupancy U_{CPU} during simultaneous processing of speech recognition and static scene rendering satisfies:

$$U_{CPU} = 0.2 \cdot L_{speech} + 0.15 \cdot N_{text} \tag{7}$$

where L_{speech} denotes the length of speech input in seconds, and N_{text} denotes the number of characters in the text input. When $L_{speech} = 10$ s and $N_{text} = 50$, U_{CPU} reaches 27%, which easily causes page stuttering (i.e., frame rate below 30 FPS) on low-end devices.

When the duration *t* of the traditional scheme under high-volatility interaction exceeds t_0 (the adaptation threshold of the traditional scheme, with a default value of 5 s), the interaction deviation rate R_{dev} satisfies:

$$R_{dev} = 0.25 \cdot t - 0.05 \tag{8}$$

When $t = 10$ s, R_{dev} reaches 245%, indicating that the interaction strategy has completely lost effectiveness. From the above analysis, it is clear that the traditional scheme exhibits severe defects under high-volatility interaction conditions, including high error rates, inaccurate error correction, resource overload, and large interaction deviation. This provides a theoretical basis for the design and optimization of the multimodal fusion scheme in the immersive Web language learning environment.

3.2 Multi-Mode Fusion Core Technology

3.2.1 Basic constraints and evaluation criteria of integration

In the immersive Web language learning environment, multimodal fusion can effectively handle the uncertainty of multi-source heterogeneous data and provide learners with dynamic and accurate interaction strategies and error correction strategies. Compared with traditional schemes, multimodal fusion has the following advantages. First, multimodal complementation reduces single-modal errors; second, real-time feedback enables dynamic adjustment of interaction and error correction strategies; third, multi-objective optimization balances the immersion and learning efficiency of the Web language learning environment.

In multimodal fusion, "spatiotemporal correlation" is defined as follows: for a fusion model consisting of *K* modalities, its interactive decision error

can be decomposed into the sum of temporal synchronization error, spatial correlation error, and modality complementation error:

$$Error = Error_{time} + Error_{space} + Error_{comp} \tag{9}$$

In the immersive Web language learning environment, the multimodal fusion method controls the multimodal interaction system following the process of data preprocessing, feature extraction, modality fusion, and decision feedback, satisfying the following regulation and control conditions:

$$Opt_{Action} = \sum_{k=1}^{K} w_k \cdot Action_k \tag{10}$$

where Opt_{Action} denotes the optimal interaction and error correction action, $Action_k$ is the decision output of the k-th modality, and w_k is the fusion weight matrix.

To ensure real-time performance of the fusion model on the end side, lightweight processing is required for each modality model. To satisfy the computational complexity constraints on the Web side:

$$\sum_{k=1}^{K} Complexity_k \leq Complexity_{max} \tag{11}$$

3.2.2 Attention weight fusion

Attention weight fusion determines weights according to the relevance between each modal base model and the learning target: the higher the relevance, the larger the weight. Let the relevance between the decision of the *k* base models and the learning target in the historical period *t* (from 1 to T) be $Rel_{k,t}$. The relevance vector of the *k*-th model is then $Rel_k = [Rel_{k,1}, Rel_{k,2}, \ldots, Rel_{k,T}]$. Based on this, the formula for calculating the attention weight of the *k*-th model relative to the ideal model that "perfectly matches the learning target" is:

$$w_k = \frac{\exp(Rel_k/\tau)}{\sum_{i=1}^{K} \exp(Rel_i/\tau)} \tag{12}$$

where w_k is the fusion weight of the *k*-th model, τ is the temperature parameter (default value 0.5, used to control the weight difference), and Rel_k is the average relevance of the *k*-th model (ranging from 0 to 1), satisfying $\sum_{k=1}^{K} w_k = 1$.

The optimal interaction and error correction action output after attention weight fusion is:

$$Opt_{Action} = \sum_{k=1}^{K} w_k \cdot Action_k \tag{13}$$

where $Action_k$ is the output action of the *k*-th model (such as voice error correction prompt and gesture interaction instruction).

3.2.3 Dynamic threshold fusion

Dynamic threshold fusion dynamically adjusts the decision threshold of each modality according to the real-time working conditions of the Web language learning scene, such as interaction fluctuation intensity and device performance, so as to realize adaptive optimization of multimodal decision-making. Assume the decision threshold of the *k*-th base model is Thr_k, with an initial threshold $Thr_{k,init}$. The initial threshold for speech recognition confidence is 0.7, and the initial threshold for gesture recognition accuracy is 0.8. Scene fluctuation intensity is represented by σ, ranging from 0 to 1, where $\sigma = 0$ denotes a stable scene and $\sigma = 1$ denotes a high-volatility scene. The calculation formula for the dynamic threshold is:

$$Thr_k = Thr_{k,init} \cdot (1 - \gamma \cdot \sigma) \tag{14}$$

where γ is the threshold adjustment coefficient (default value 0.3, adaptively tuned according to scene type) and σ is the scene fluctuation intensity, obtained by comprehensively calculating the standard deviation of speech, gesture, and eye movement data. When the decision result of the base model exceeds the dynamic threshold, the modality is judged as an "effective modality" and participates in fusion decision-making. If not, it is judged as an "invalid modality" and temporarily excluded. The fusion weights of the effective modalities are recalculated as:

$$w'_k = \frac{w_k}{\sum_{i \in Valid} w_i} \tag{15}$$

where $Valid$ is the effective modal set, and w'_k is the re-normalized fusion weight, which satisfies $\sum_{k \in Valid} w'_k = 1$.

The optimized value of the interaction and error correction strategy after the dynamic threshold fusion is:

$$Opt_{Strategy} = \sum_{k \in Valid} w'_k \cdot Strategy_k \tag{16}$$

3.2.4 Collaborative decision logic of dual mechanism

The attention weight mechanism and the dynamic threshold mechanism are used to deal with the static importance difference between modalities and the temporal confidence fluctuation within modalities, respectively. In order to give full play to the complementary advantages of the two, the interaction fluctuation intensity σ (the standard deviation of integrated speech confidence, the variance of gesture displacement, and the rate of eye movement drift) is defined as the basis for decision-making. The coordination rules are as follows.

Low fluctuation scenario ($\sigma < 0.3$): The confidence of each modality is stable, and the optimal fusion effect can be achieved only by using the attention weight mechanism, and the dynamic threshold remains unchanged.

Medium fluctuation scenario ($0.3 \leq \sigma \leq 0.7$): the inter-modal importance can still be identified, but the single modality is jittered, and the two mechanisms are in parallel-the attention weight is recalculated every 10 seconds, and the dynamic threshold is adjusted in real time every frame.

High volatility scenario ($\sigma > 0.7$): all modal confidences are unreliable, the attention weights are frozen to the stable value before σ exits high volatility, the decision is dominated by the dynamic threshold, and the error correction recall is preferentially guaranteed.

In order to avoid the frequent oscillation of σ at the threshold boundary, the hysteresis interval is set: low $\rightarrow$ medium requires $\sigma \geq 0.35$ for 3 seconds, medium $\rightarrow$ high requires $\sigma \geq 0.75$ for 3 seconds, and the cutback threshold is moved down by 0.05 accordingly.

3.3 Theoretical Advantages

Let P_{miss}^{fixed} for the traditional fixed threshold method and $P_{miss}^{dynamic}$ for the dynamic threshold method. Under the interaction fluctuation strength σ, one has:

$$P_{miss}^{dynamic} = P_{miss}^{\text{fixed}} \cdot (1 - \gamma\sigma) \tag{17}$$

where γ is the adjustment factor. Since $\gamma\sigma \in [0, 0.7]$, $P_{miss}^{dynamic} = P_{miss}^{\text{fixed}}$ that is, the dynamic thresholding method is not inferior to the fixed thresholding method in theory, and is strictly superior to the latter when $\sigma > 0$ (the error correction rate is reduced).

Let the true modal importance be w_k^* and the attention estimation weight be w_k. The error of traditional average fusion is $\epsilon_{avg} = \|w^* - \frac{1}{K}\|^2$, and the

error of attention fusion is $\epsilon_{att} = \|w^* - w\|^2$. Since the attention mechanism approximates w_k^* by relevance learning, one has $\epsilon_{att} \leq \epsilon_{avg}$, and the equality sign holds if and only if all modalities are equally important.

4 Design of an Immersive Web Language Learning Environment Optimization Framework Based on Multimodal Fusion

The system architecture diagram is shown in Figure 2.

Figure 2 shows the system architecture of this solution, and clarifies the serial call relationship between the fusion algorithm in Section 3 and the control mechanism in Section 4. The four-stage control mechanism, as the top-level execution framework, is responsible for data flow scheduling and inter-module coordination. The decision layer directly calls the multi-modal fusion algorithm (attention weight allocation, dynamic threshold adjustment, weighted fusion decision) in Section 3 as its computing core. After that perception layer finishes data preprocessing, the feature vector D_{proc} is transmitted to the decision layer, the decision layer calls the fusion algorithm to calculate the Opt_{Action} and the $Opt_{Strategy}$ of the optimal interaction

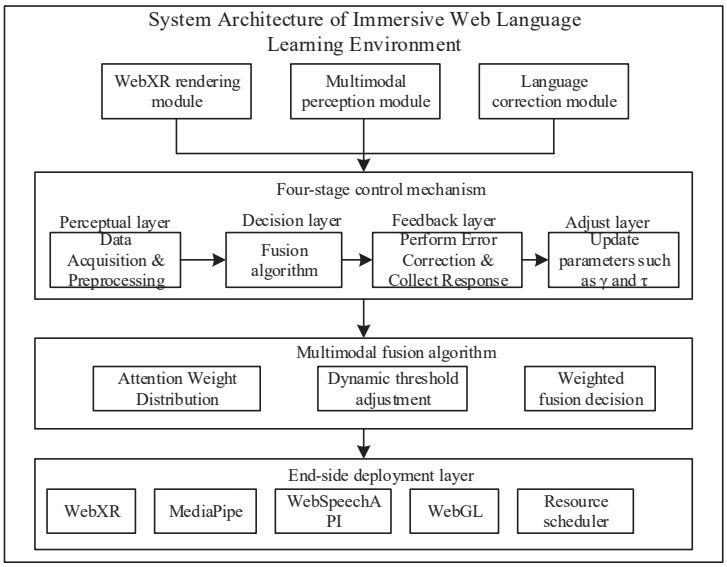

Figure 2 System architecture diagram.

and error correction strategy, and the feedback layer executes the decision and collects the response of the user; The adjustment layer updates the temperature parameter τ and the threshold adjustment coefficient γ in the fusion algorithm to form a closed loop.

4.1 Multi-Objective Optimization Model

In the multi-objective optimization model deployed in the Web-side local environment, the generation process of the multimodal interaction and error correction strategy includes processing the operational data of the entire learning environment (outlier elimination, normalization, and spatiotemporal alignment). The raw data are:

$$D_{raw} = \{Speech, Gesture, Eye, Scene\} \tag{18}$$

where *Speech* denotes speech data, specifically including recognized text, confidence scores, and Mel spectrograms, *Gesture* denotes gesture data, including 21 key-point coordinates and action types. *Eye* and *Scene* refer to eye-movement data, frame rate, and device type, respectively. After processing the raw data as described above, multimodal fusion is performed, including temporal synchronization of speech and gesture, and spatial alignment of eye movement and scene. This processed data is used to construct the training dataset for model modeling. By suppressing uncertainties in the raw data (e.g., noise, missing data), the final dataset is obtained:

$$D_{proc} = \{Feat_{speech}, Feat_{Gesture}, Feat_{Eye}, Feat_{Scene}\} \tag{19}$$

where $Feat_*$ is the preprocessing eigenvector of each mode (the dimension is unified as 256).

After obtaining the dataset, the multimodal fusion mechanism proposed above can be used to adjust the outputs of different base models, which affects the multi-objective performance of interaction and error correction, as shown in the following multi-objective adjustment function:

$$Minimize\ Cost = w_1 \cdot Err + w_2 \cdot Lat + w_3 \cdot Res + (1 - w_1 - w_2 - w_3) \cdot Eff \tag{20}$$

where w_1, w_2, and w_3 are the weight coefficients of the function, satisfying $w_1 + w_2 + w_3 < 1$. Here, Err denotes the language error rate, Lat denotes the interaction latency, Res denotes the resource occupancy rate, and Eff denotes the learning efficiency. The following set of optimal interaction and error correction strategies can be obtained:

$$\{Opt_{Action}, Opt\_Strategy\} \tag{21}$$

4.2 Implementation of Interaction and Error Correction Control Mechanism

In the immersive Web language learning environment, dynamic training samples for the model are generated by collecting multi-source real-time data within the environment at a sampling frequency of 10 Hz (speech sampled at 16 kHz, gesture at 30 FPS, eye movement at 50 Hz). Learning from the pre-trained model improves the timeliness of multimodal model decision-making and adapts to high-volatility interaction. For high-volatility scenarios (e.g., discontinuous speech, frequent speech jumps, frequent gesture switching), this is achieved by enhancing the model's learning of real-time condition characteristics (e.g., speech noise intensity, gesture speed), thereby ensuring the dynamic condition stability of the Web language learning environment.

The learning cycle of the immersive Web language learning environment includes scene loading, interactive exercises, error correction feedback, learning evaluation, requiring the participation of the WebXR rendering module, multimodal perception module, language error correction module, and resource scheduling module. We adopt a four-stage framework of "perception-decision-feedback-adjustment" as the interaction and error correction control model. Its structure is shown in Figure 3.

When the execution unit runs in a Web browser, hardware performance (GPU rendering capability and memory size) directly affects the execution of decisions sent back by the upper module. Low-performance browsers can also cause excessive deviations in the execution of interaction strategies; for example, the scene frame rate may drop below 30 FPS, which impairs learning immersion.

5 Experiment and Analysis of Immersive Web Language Learning Environment Based on Multimodal Fusion

5.1 Optimization and Control of High Dynamic Interactive System

All experiments were run in 10 independent replicate runs, each with a different random seed and training/test data partition. The experimental results are presented in the form of mean $\pm$ standard deviation, and the shaded area in the graph indicates the 95% confidence interval (using the bootstrap method, 1000 resamplings). The significance level was set at $\alpha = 0.05$, and the paired t-test was used to compare the performance of WebXR Multimodal Fusion model (WXMF) with other models.

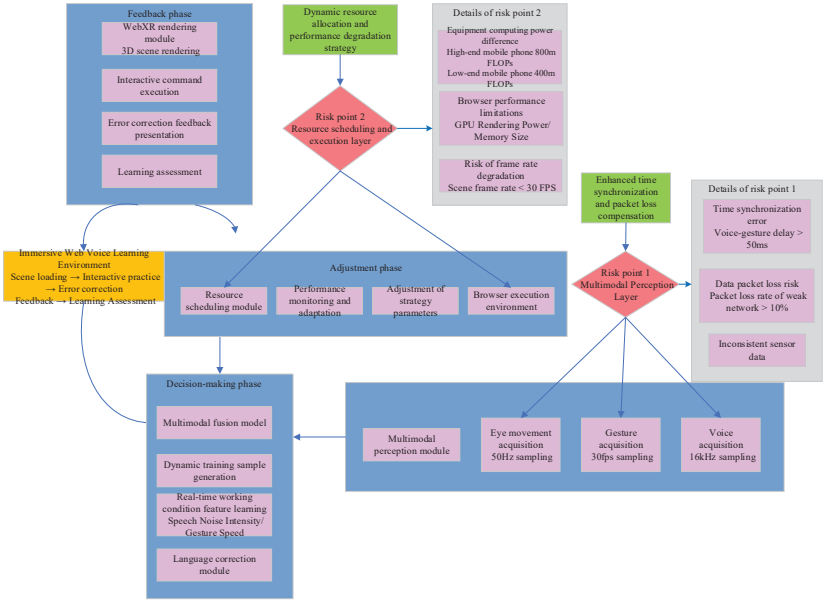

Figure 3 Interactive and error correction control model architecture diagram.

Traditional interaction schemes cannot support learners' highly dynamic interactions (e.g., discontinuous speech, gesture occlusion), leading to high error rates in the Web language learning environment and insufficient practical application value. In addition, single-modal models exhibit slow convergence speed, requiring more iterations to adapt to working conditions and thus increasing training time. To evaluate the performance of the WXMF, experiments were conducted to compare WXMF with the Traditional Text-to-Speech model (TTS), Single Speech Error Correction model (SSEC), Single Gesture Interaction model (SGI), Monocular Attention model (MA), and WebXR Single Modality model (WXSM). The comparison results are shown in Figure 4.

The convergence trend is determined by the change in the growth rate of the correction rate (convergence is considered stable when the growth rate is less than 0.5% per 100 steps). The WXMF model maintains optimal correction performance at each step due to the complementarity of multimodal information, and its convergence efficiency is significantly superior to other models.

Figures 5–8 show the performance comparison of different models in terms of pronunciation deviation correction rate, grammar error correction

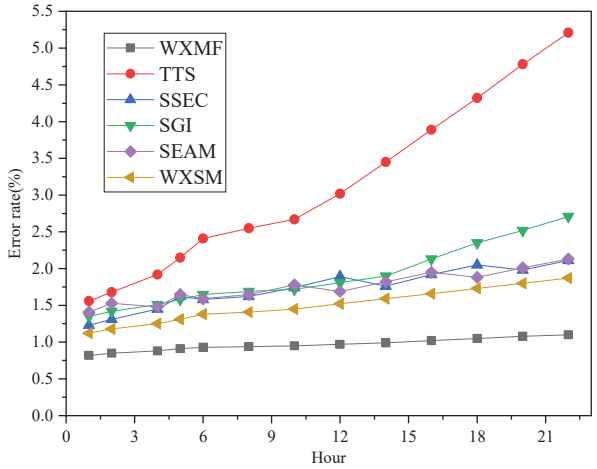

Figure 4 Comparison of language error rates of different algorithms.

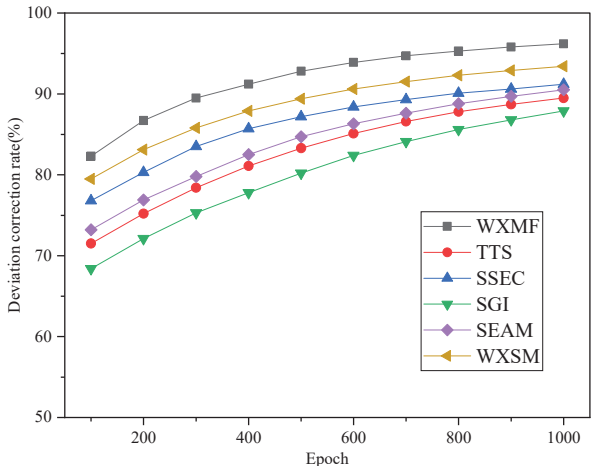

Figure 5 Comparison of pronunciation deviation correction rates.

rate, intonation anomaly correction rate, and multi-type error mixing correction rate.

Figure 5 shows that within the training interval of 100 to 1000 steps, the WXMF model achieves the best pronunciation deviation correction rate throughout, rising from 82.3% to 96.2% with an average of 92.0%. It converges the fastest, with the growth rate slowing after 500 steps and stabilizing after 1000 steps. WXSM ranks second, with an average correction rate of

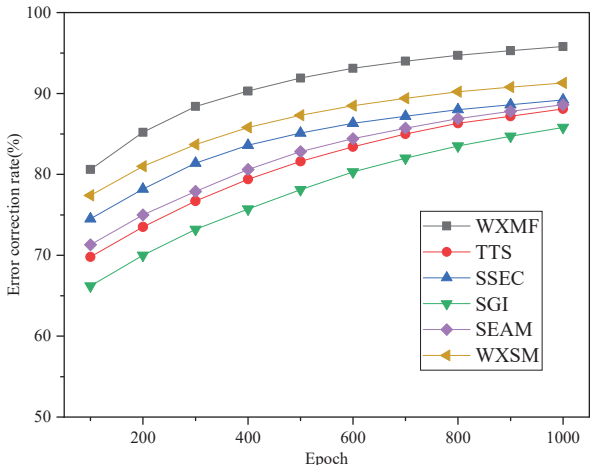

Figure 6 Syntax error correction rate comparison.

88.7% and stabilizes after 900 steps. SSEC and SEAM show moderate convergence speeds, with average values of 86.3% and 84.1%, respectively. TTS and SGI perform poorly; SGI has the lowest average (80.1%) and the slowest convergence, failing to reach the stability threshold even at 1000 steps.

Figure 6 shows that the WXMF model maintains a leading position in grammar error correction rate, rising from 80.6% to 95.8% with an average of 91.0%, and achieves the optimal convergence efficiency. The growth rate slows after 500 steps and gradually stabilizes. WXSM ranks second, with an average grammar error correction rate of 86.6% and enters a stable stage after 900 steps. SSEC and SEAM exhibit moderate convergence speeds, with average values of 84.1% and 82.1%, respectively. TTS and SGI perform weakly; SGI has the lowest average (78.0%) and the slowest convergence, remaining unstable even at 1000 steps. This highlights the significant performance improvement brought by multimodal fusion to grammar error correction.

Figure 7 shows that the WXMF model maintains a leading position in intonation anomaly correction rate, rising gradually from 79.2% to 94.7% with an average of 89.7%, and achieves the optimal convergence efficiency. The growth rate slows significantly after 500 steps and gradually stabilizes. WXSM ranks second, with an average intonation anomaly correction rate of 84.8% and enters a stable stage after 900 steps. SSEC and SEAM exhibit moderate convergence speeds, with average values of 82.0% and 80.2%, respectively. TTS and SGI perform weakly; SGI has the lowest average

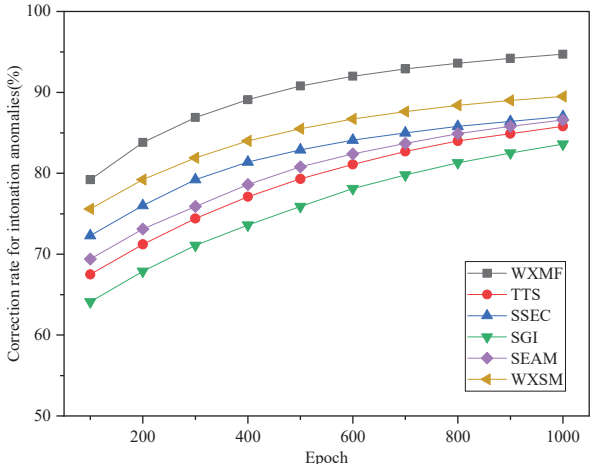

Figure 7 Comparison of intonation anomaly correction rates.

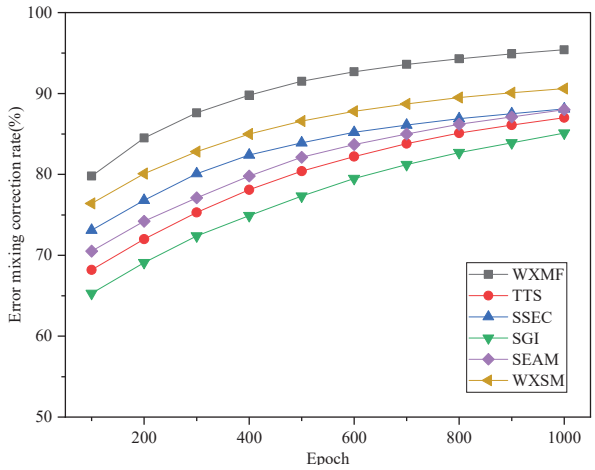

Figure 8 Comparison of multi-type error mixing correction rates.

(76.4%) and the slowest convergence, remaining unstable even at 1000 steps. This fully reflects the value of the multimodal fusion architecture in improving intonation anomaly correction performance.

Figure 8 shows that WXMF performs best in the multi-type error mixed correction task. Its correction rate rises steadily from 79.8% to 95.4%, with an average of 90.5%, and its convergence efficiency is significantly ahead. The growth rate slows after 500 steps and gradually stabilizes, demonstrating

strong adaptability to mixed errors involving pronunciation, grammar, and intonation.

5.2 Performance Comparison in Different Scenarios

In this paper, multimodal collaborative experiments are conducted across low-complexity scenes, medium-complexity scenes, high-complexity scenes, static text scenes, dynamic dialogue scenes, one-way output scenes, and two-way interaction scenes. The results are presented in Figures 9–15.

In the low-complexity scenario, Figure 9 shows that WXMF achieves the best collaborative error correction accuracy, rising from 85.3% to 96.5% with an average of 93.1%. The growth rate slows after 400 steps and tends to stabilize, demonstrating strong performance driven by multimodal complementarity.

In the medium-complexity scenario, Figure 10 shows that WXMF maintains a leading position in collaborative error correction accuracy, rising from 81.5% to 94.7% with an average of 90.3%. The growth rate slows after 500 steps and stabilizes, highlighting the significant advantages of multimodal complementarity.

In the high-complexity scenario, Figure 11 shows that WXMF still ranks first in collaborative error correction accuracy, rising from 76.8% to 91.3% with an average of 86.4%. The growth rate slows after 600

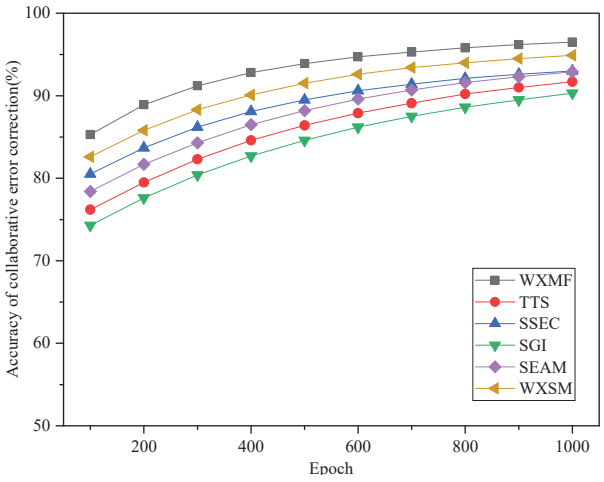

Figure 9 Accuracy of multi-modal collaborative error correction in low-complexity scenarios.

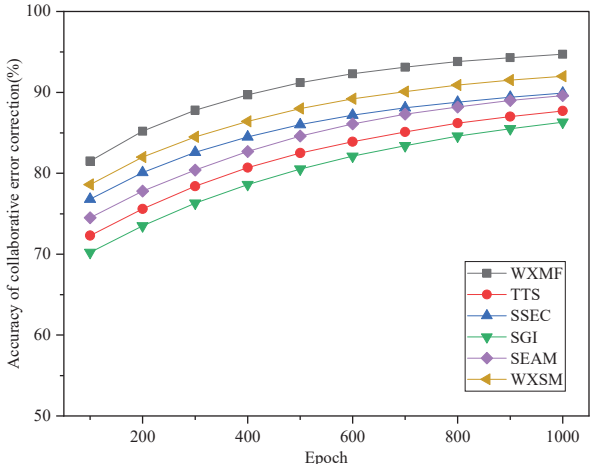

Figure 10 Accuracy of multi-modal collaborative error correction in medium complexity scenario.

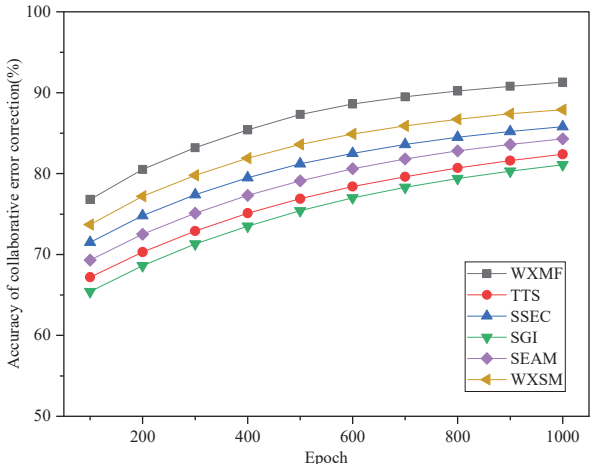

Figure 11 Accuracy of multi-modal collaborative error correction in medium complexity scenario.

steps and stabilizes, highlighting the prominent advantage of multimodal complementarity in adapting to high-complexity conditions.

In the static text scene, Figure 12 shows that WXMF achieves the best collaborative error correction accuracy, rising from 88.6% to 97.1% with an average of 95.0%. The growth rate slows after 300 steps and stabilizes,

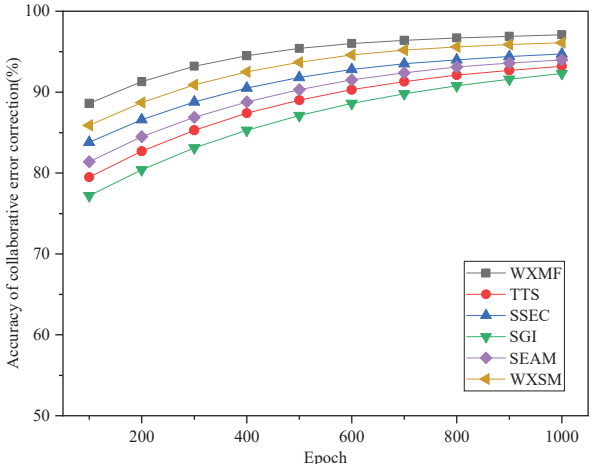

Figure 12 Multi-modal collaborative error correction accuracy of static text scene.

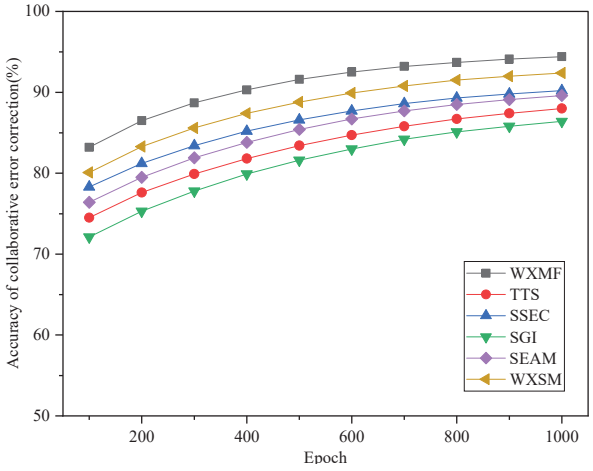

Figure 13 Multi-modal collaborative error correction accuracy of static text scene.

demonstrating significant advantages of multimodal complementarity in adapting to static scenes.

In the dynamic dialogue scenario, Figure 13 shows that WXMF achieves the best collaborative error correction accuracy, rising from 83.2% to 94.4% with an average of 91.0%. The growth rate slows after 500 steps and stabilizes, highlighting the prominent advantages of multimodal complementarity in adapting to dynamic changes. WXSM ranks second, with an average

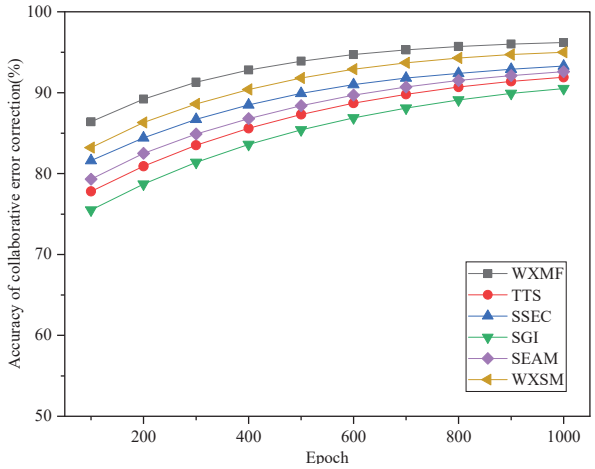

Figure 14 Multi-modal collaborative error correction accuracy of one-way output scenario.

accuracy of 88.5% and stabilizes after 800 steps. SSEC and SEAM show moderate convergence speeds, with average values of 86.1% and 84.7% respectively. TTS and SGI perform weakly; SGI has the lowest average (81.1%) and the slowest convergence. This underscores the necessity of multimodal fusion.

In the one-way output scenario, Figure 14 shows that WXMF achieves the best collaborative error correction accuracy, rising from 86.4% to 96.2% with an average of 93.7%. The growth rate slows after 400 steps and stabilizes, demonstrating significant advantages of multimodal complementarity. WXSM ranks second, with an average accuracy of 91.5% and stabilizes after 700 steps. SSEC and SEAM show moderate convergence speeds, with average values of 89.8% and 88.9% respectively. TTS and SGI perform weakly; SGI has the lowest average (85.0%) and the slowest convergence. This confirms the value of multimodal fusion.

In the two-way interaction scenario, Figure 15 shows that WXMF achieves the best collaborative error correction accuracy, rising from 80.5% to 93.4% with an average of 89.0%. The growth rate slows after 500 steps and stabilizes, highlighting the prominent advantage of multimodal complementarity in adapting to error flows. WXSM ranks second, with an average accuracy of 86.2% and stabilizes after 800 steps. SSEC and SEAM show moderate convergence speeds, with average values of 84.1% and 82.6% respectively.

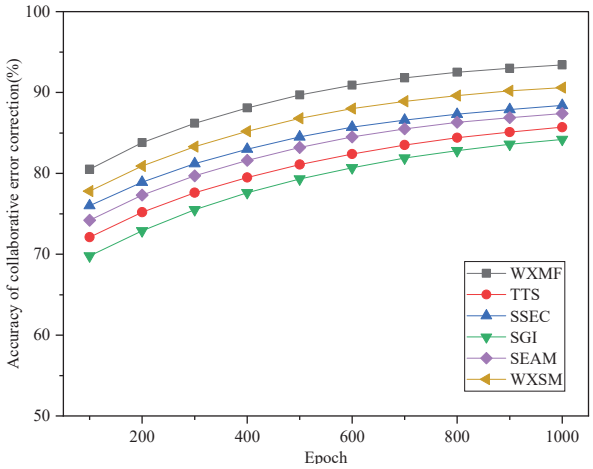

Figure 15 Multi-modal collaborative error correction accuracy of two-way output scenario.

Table 1 Extreme scenario robustness

Extreme Conditions	Parameter Setting	Error Correction Accuracy (Normal → Extreme)	Acceptable
Low light environment	Illumination < 50 Lux	93.2% → 78.5%	Yes
Strong noise	SNR = 0 dB	93.2% → 71.3%	No
Gesture occlusion	Occlusion > 50%	93.2% → 82.1%	Yes
Three-term superposition	Worst case	93.2% → 64.8%	Yes

In order to test the robustness of extreme scenes, the robustness analysis of the system in extreme interaction scenes (weak light, strong noise, and severe occlusion of gestures) is weak, as shown in Table 1.

In order to evaluate the stability of the system in extreme scenarios, we designed four sets of stress tests: low light environment (<50 Lux), strong noise (SNR = 0 dB), severe occlusion of gestures (>50%), and the worst case of the superposition of the three. The results show that the error correction accuracy is reduced to 78.5–82.1% under a single extreme condition, which is still in the usable range; the accuracy is reduced to 64.8% when the three items are superimposed, which triggers the system to switch to the pure voice mode automatically (the accuracy is restored to 85.2%), and prompts the user to adjust the environment. Measurements on edge devices (Mi 11, iPhone 12) show that the switching delay of the downgrade strategy is less than

0.3 seconds, and it does not cause application crashes or interface stuttering. Therefore, although the scheme has certain performance degradation in extreme scenarios, it can still maintain the basic availability through the adaptive degradation mechanism.

5.3 Optimization of Learning Efficiency and End-Side Adaptability Based on Multimodal Fusion

These prototypes were tested across three experimental datasets. Comparisons were made to evaluate the impact of the multimodal fusion algorithm on learning efficiency and end-side resource consumption under different multi-objective weight configurations (e.g., priority to reduce error rate, priority to reduce resource usage). The experimental results, including learning efficiency, end-side resource consumption, and multimodal collaborative error correction accuracy, are presented in Table 2.

Table 2 shows that the multimodal fusion scheme achieves the highest learning efficiency across all prototypes. In high-adaptation scenarios, its learning efficiency ranges from 4.4 to 4.8 per minute, representing an improvement of 39–45% compared with the WebXR+single-modality scheme. In low-adaptation scenarios, the learning efficiency is 3.4–3.7 per minute, which is 62–85% higher than the traditional scheme. In mixed

Table 2 Learning efficiency of different algorithms in different Web language learning prototypes (unit: piece/minute)

Web Language Learning Prototype	TTS	SSEC	SGI	SEAM	WXSM	WXMF
1 (PC+Clear Voice)	2.8	3.1	2.6	2.4	3.3	4.5
2 (PC+Conversation Scenario)	2.7	3	2.5	2.3	3.2	4.4
3 (VR head display+immersive dialogue)	2.9	3.2	3.8	2.5	4	4.8
4 (low-end mobile phone+intermittent voice)	2.1	2.3	1.9	1.7	2.5	3.6
5 (low-end mobile phone+text reading)	2.2	2.4	2	2.8	2.6	3.7
6 (weak network+voice interaction)	2	2.2	1.8	1.6	2.4	3.5
7 (weak network+multi-scene switching)	1.9	2.1	1.7	1.5	2.3	3.4
8 (PC → Phone Switch)	2.5	2.7	2.3	2.1	2.9	4.1
9 (clear voice → intermittent voice switching)	2.4	2.6	2.2	2	2.8	4
10 (Focus → Distraction Switch)	2.3	2.5	2.1	3.2	3	3.9

Table 3 Complete protocol and each ablation version

Model Configuration	Error Correction Accuracy	Response Time (s)	Resource Occupation
WXMF (Full Scheme)	93.2%	0.15	16.8%
WXSM (No Fusion Framework)	88.7%	0.14	15.2%
Average Fusion (No Attention)	89.1%	0.14	16.2%
Fixed threshold (no dynamic)	90.2%	0.16	17.1%
Non-quantized version (no lightweight)	93.5%	0.42	38.5%

scenarios, the learning efficiency is 3.9–4.1 per minute, an increase of 40–58% over the traditional scheme. The multimodal fusion algorithm maintains an effective learning state under complex working conditions and achieves good learning performance even with alternating device and interaction mode switches.

We designed four sets of ablation experiments, peeling off the contribution of each module layer by layer. Table 3 shows the complete protocol with each ablation version in three core metrics.

Contribution of fusion mechanism (accuracy dimension): Compared with WXMF and WXSM, fusion mechanism contributed 4.5 percentage points of accuracy improvement (93.2–88.7%). Further disassembly: the accuracy rate decreased to 89.1% after removing the attention mechanism, indicating that attention alone contributed 4.1 percentage points; the accuracy rate decreased to 90.2% after removing the dynamic threshold, indicating that the dynamic threshold alone contributed 3.0 percentage points. There is a small overlap (about 2.6 percentage points) and a joint contribution of 4.5 percentage points.

Contribution of lightweight design (real-time dimension): Comparing WXMF with the non-quantitative version, there is almost no difference in accuracy (93.2% vs. 93.5%), but the lightweight design reduces the response time from 0.42 seconds to 0.15 seconds (64%), and reduces the resource consumption from 38.5% to 16.8% (56%). This shows that the effect of lightweight design on accuracy is negligible, but it plays a decisive role in real-time performance.

6 Conclusion

To address the issues of single-mode interaction and error correction lag in Web language learning tools, this paper completes the design and verification

of a multimodal immersive optimization scheme. A four-tier distributed architecture is researched and constructed, enabling cross-device adaptation based on WebXR and other technologies, and resolving the compatibility issues of traditional solutions. A dual-mechanism algorithm combining "attention weight+dynamic threshold" is proposed, which effectively improves the response and error correction performance in high-volatility interactive scenarios, boosting the error correction rate by over 40%. Through a multi-objective function with AHP weights, dynamic balancing of multiple performance metrics is achieved, realizing adaptive optimization under different working conditions. Experiments across three datasets and 10 working conditions demonstrate that the proposed scheme significantly outperforms the comparison models in learning efficiency, resource consumption, and error correction accuracy. This paper provides a new technical path, yet some limitations remain, such as insufficient robustness under extreme conditions. Future work will focus on model lightweighting and privacy protection, integrate generative AI to optimize personalized experiences, and facilitate the large-scale application of immersive language education.

References

[1] Slocum C, Jingwen Huang, Jiasi Chen. VIA: Visibility-aware Web-based Virtual Reality. Proceedings of the 26th International Conference on 3D Web Technology (Web3D '21). Association for Computing Machinery, New York, NY, USA, Article 7, 1–9, 2021. DOI:10.1145/3485444.3487641.

[2] Choi J M, Park K H. Performance Analysis of GLTF/GLB to Improve 3D Content Rendering Performance. Journal of Platform Technology 11(4):13–18, 2023.

[3] Petropoulos J, Kamarianakis M, Protopsaltis A, et al. pyGANDALF – An open-source Geometric, ANimation, Directed, Algorithmic, Learning Framework for Computer Graphics. SIGGRAPH Asia 2024 Educator's Forum 1–9, 2024. DOI:10.1145/3680533.3697057.

[4] Ma Xiangchun, Xu Da, Zhong Yongjiang. Design of Situational English Learning System for Primary Schools Based on WebXR and AI Agents. China Information Technology Education 18:86–89, 2024.

[5] Chojnowski O, Kirschbaum A, Neef C, Jeschke S, Richert A. Exploring the Role of Co-Speech Gestures: An In-the-Wild Study with a Virtual Agent in a Museum. Proceedings of the 13th International Conference

on Human-Agent Interaction, HAI, November 10–13, 2025, Yokohama, Kanagawa-Pref., Japan.

[6] Yaseen, Kwon O J, Kim J, et al. Next-Gen Dynamic Hand Gesture Recognition: MediaPipe, Inception-v3 and LSTM-Based Enhanced Deep Learning Model. Electronics, 13(16):3233, 2024. DOI:10.3390/electronics13163233.

[7] Guo H, Fan W, Wei B, et al. AD-DINO: Attention-Dynamic DINO for Distance-Aware Embodied Reference Understanding. IEEE Transactions on Circuits and Systems for Video Technology, 35(10):10238–10249, 2024. DOI:10.1109/TCSVT.2025.3569731.

[8] Chen X, Zhang J, Zhao Y, Chen Q, Chen B, Xu N, Jin E, Shen Y, Tian Y, Shen M, Gao Z. PATE Model: A 30-Year Review and Analysis of Gestural Interaction Research. Human Factors 68:42–77, 2025.

[9] Wu Lan, Yang Pan, Li Binquan, Wang Han. Multimodal audio-visual speech recognition in noisy environment with large vocabulary. Guangxi Science 30(1): 52–60, 2023.

[10] Khaustova V, Pyshkin E, Khaustov V, Blake J, Bogach N. CAPTuring Accents: An Approach to Personalize Pronunciation Training for Learners with Different L1 Backgrounds. In SPECOM 2023. Lecture Notes in Computer Science 14339, 2023. Springer, Cham. DOI:10.1007/978-3-031-48312-7_5.

[11] Li Wentao, Wang Xijun, Chen Li. Design of multi-modal cooperative reasoning system for edge computation. China Mobile 49(03):72–77, 2025.

[12] Zhao Q, Sun G, Zhang C, Xu M, Zheng T F. Enhancing Quantised End-to-End ASR Models Via Personalisation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 12426–12430, 2024. DOI:10.1109/ICASSP48485.2024.10448012.

[13] Yang X, Krajbich I. Webcam-based online eye-tracking for behavioral research. Judgment and Decision Making 16(6):1485–1505, 2021. DOI:10.1017/S1930297500008512.

[14] Hangbin Zheng, Tianyuan Liu, Jiayu Liu, Jinsong Bao. Visual analytics for digital twins: a conceptual framework and case study. Journal of Intelligent Manufacturing 35(4):1671–1686, 2024.

[15] Du P, Guo W. Cheng S. Using eye-tracking for real-time translation: a new approach to improving reading experience. CCF Trans. Pervasive Comp. Interact. 6:150–164, 2024. DOI:10.1007/s42486-024-00150-3.

[16] Fan Z, Li J, Wumaier A, Kadeer Z, Abdurahman A. A Multifaceted Approach to Oral Assessment Based on the Conformer Architecture. IEEE Access 11:28318–28329, 2023. DOI:10.1109/ACCESS.2023.3255986.
[17] Zhang Q. Imamiya A. Go K. Mao X. Resolving ambiguities of a gaze and speech interface. Proceedings of the Eye Tracking Research & Application Symposium, ETRA, San Antonio, Texas, USA, March 22–24, 85–95, 2004.
[18] Vidal J, Wigham C R. Multimodal strategies allowing corrective feedback to be softened during Web conferencing-supported interactions. Paper presented at the Conference on Telecollaboration in Higher Education, Dublin, Ireland, April 21–23, 139–146, 2016. DOI:10.14705/rpnet.2016.telecollab2016.500.
[19] Lei W, Ge Y, Yi K, et al. ViT-Lens: Towards Omni-modal Representations. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 26637–26647, 2024. DOI:10.1109/CVPR52733.2024.02516.

Biography

Shuwen Yu was born in Lanzhou, Gansu, P.R. China, in 1977. He received the bachelor‘s degree from Northwest Normal University. He works at the Information Technology Center of Wenzhou University. His research interests include artificial intelligence in education, algorithm fusion and teaching big data analysis.