Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 21 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i21.58853

PAPER

Design of a Virtual Reality-Supported Immersive English Learning Environment and Interaction Behavior Analysis

Lanlan Fang[1], Xiujuan Wang[2] (✉), Liang Zhang[1]

[1]Hunan Mechanical & Electrical Polytechnic, Changsha, China

[2]Hunan Railway Professional Technology College, Zhuzhou, China

p121690@siswa.ukm.edu.my

ABSTRACT

To address limitations in current virtual reality (VR)-based English learning environments—specifically the insufficient accuracy of real-time speech recognition, restricted interactive feedback, and the lack of advanced behavioral analysis capabilities—a comprehensive solution integrating advanced speech recognition techniques with immersive system design was developed. The study is organized into two major components. First, an end-to-end speech recognition model tailored for immersive VR environments was designed and implemented. The model integrates a low-rank feed-forward module with a probabilistic sparse self-attention mechanism, substantially reducing computational complexity while maintaining real-time processing performance on consumer-grade VR devices. Furthermore, a dual-path decoder architecture combining Connectionist Temporal Classification (CTC) with an attention-based mechanism was employed. This design simultaneously satisfies the low-latency requirements of streaming recognition and leverages rescoring in non-streaming modes to achieve higher accuracy, thereby improving robustness and precision in recognizing continuous, natural English speech. Second, building upon this recognition model, a fully integrated VR-based English learning system was designed. The system generates contextualized language tasks, supports multimodal human-computer interaction, and comprehensively records both speech and behavioral data throughout the learning process. This approach not only provides learners with a highly immersive and interactive environment featuring immediate pronunciation feedback and personalized training but also establishes a robust data foundation and technical platform for in-depth analysis of interaction behaviors and mechanisms of language acquisition. These contributions highlight significant theoretical value and practical potential for advancing intelligent education.

KEYWORDS

virtual reality (VR), English language learning, end-to-end speech recognition, conformer, interaction behavior analysis, immersive learning environment

Fang, L., Wang, X., Zhang, L. (2025). Design of a Virtual Reality-Supported Immersive English Learning Environment and Interaction Behavior Analysis. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(21), pp. 184–198. https://doi.org/10.3991/ijim.v19i21.58853

Article submitted 2025-07-12. Revision uploaded 2025-09-05. Final acceptance 2025-09-07.

1 INTRODUCTION

With the rapid advancement of information technologies, virtual reality (VR) [1–3] has provided unprecedented immersive, contextualized, and interactive experiences for language learning, and has become a major research direction in educational technology. In the process of learning English as a second language, authentic and natural language environments are considered critical for enhancing learners' listening and speaking skills as well as their communicative confidence. VR technology enables the simulation of realistic language-use scenarios [4–6], such as international conferences, social conversations, and business negotiations, thereby allowing learners to repeatedly practice and engage in immediate interactions within environments that approximate real-world contexts. Nevertheless, most current VR-based language learning systems [7, 8] remain dependent on pre-scripted content and limited feedback. They lack the ability to perform real-time recognition and in-depth analysis of learners' continuous and spontaneous speech, making it difficult to deliver high-quality, personalized, and adaptive language training.

Although speech recognition technologies [9, 10] have achieved remarkable progress in recent years, significant challenges remain in effectively integrating them into VR-supported English learning environments. Traditional speech recognition approaches [11–13], typically designed for general-purpose applications, have limited capacity to address the unique conditions of VR contexts, which are often characterized by acoustic noise, non-stationary input, multi-user interaction, and extended speech sequences. Furthermore, many existing systems [14–16] employ either non-end-to-end or partially streaming architectures, which fail to simultaneously meet the requirements of high accuracy and low latency, thereby constraining their applicability in real-time interactive learning. At the model level, current methods [17] often suffer from excessive computational complexity and high memory demands, limiting deployment on consumer-grade VR devices. In addition, insufficient modeling of educationally relevant features, such as pronunciation errors and pragmatic behaviors, has hindered the ability of these systems to support deeper analyses of learning behaviors and instructional interventions.

Building upon these challenges, the present study focuses on the design of a VR-supported immersive English learning environment and the analysis of interaction behaviors, with the research framework consisting of two major components. First, an end-to-end speech recognition model was designed and implemented for VR-based immersive English learning. The model integrates a low-rank feed-forward module, a probabilistic sparse self-attention mechanism, and a dual-path decoder combining Connectionist Temporal Classification (CTC) with an attention-based mechanism. This configuration ensures streaming capability and real-time responsiveness, while substantially improving the accuracy and robustness of recognizing long-sequence English speech. Immediate feedback on pronunciation and enhanced comprehension support are thereby enabled for learners. Second, a holistic VR English learning system framework was constructed, integrating speech recognition, behavior tracking, and multimodal interaction. The system incorporates contextualized task generation, real-time speech-based interaction, and comprehensive process recording, thereby supporting diverse learning modes and evaluation mechanisms. This design provides a complete technological platform for investigating language acquisition mechanisms and interaction behaviors in immersive environments.

2 SYSTEM DESIGN

2.1 End-to-end speech recognition

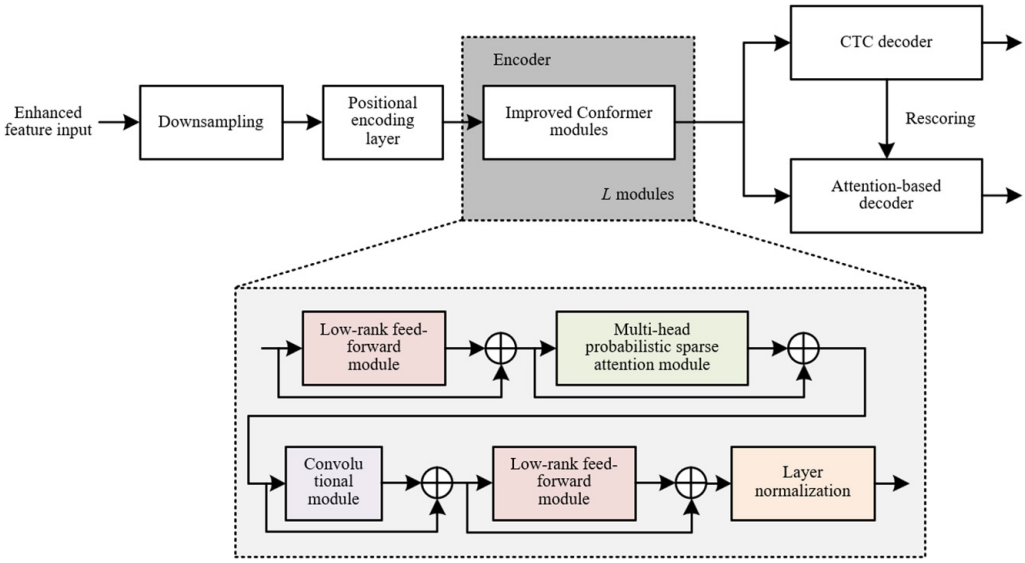

Fig. 1. Architecture of the end-to-end speech recognition model for a VR-supported immersive English learning environment

To meet the requirements of real-time interaction and behavior analysis in a VR-supported immersive English learning environment, an end-to-end speech recognition model was constructed. The model first performs speech enhancement on noisy audio collected from the VR environment, followed by the extraction of acoustic features suited to the characteristics of English speech. Because speech input in VR contexts often consists of long sequences, the excessive number of frames can substantially increase computational load. Moreover, continuous phonemes and words in English are typically represented by multiple frames. To address this issue, a downsampling mechanism was introduced to merge redundant frames and compress sequence length, thereby improving processing efficiency while preserving the integrity of acoustic information. The downsampled features are positionally encoded and then passed into a Conformer encoder composed of multiple stacked Conformer modules. Each module integrates a low-rank feed-forward network, a multi-head probabilistic sparse attention mechanism, and a convolutional module, enabling the effective capture of both long-range dependencies and local contextual features in English speech. The encoder output is directed to two parallel decoding branches: the CTC decoder and the attention-based decoder. The CTC branch supports streaming recognition, ensuring real-time dialogue and immediate feedback in VR environments, whereas the attention-based decoder performs non-streaming decoding for high-accuracy offline processing and analysis. Finally, an attention-based rescoring mechanism is employed to fuse the outputs of both decoders. This integration not only enhances recognition accuracy for English speech but also provides a reliable technical foundation for pronunciation evaluation and interaction behavior analysis. The overall model architecture is illustrated in Figure 1, and the key modules of the model are described in detail below.

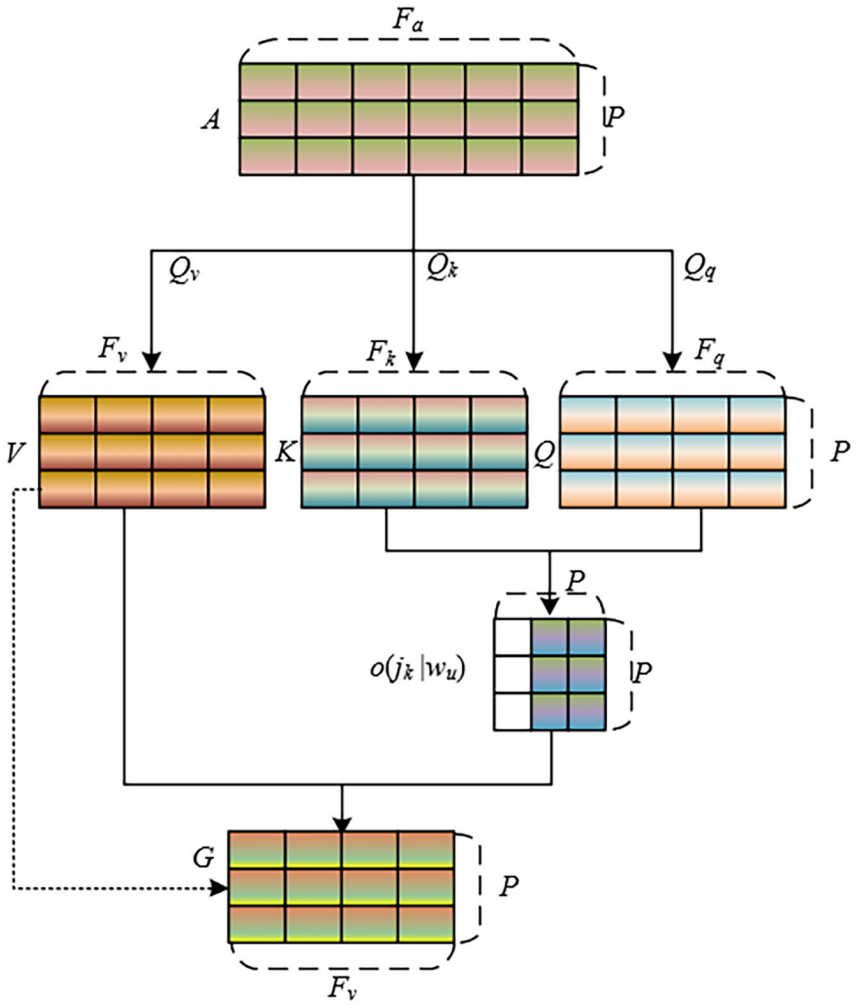

Fig. 2. Schematic illustration of the probabilistic sparse self-attention mechanism

In VR contexts, learners typically engage in continuous and natural conversational practice, resulting in long speech input sequences. When a standard self-attention mechanism is applied, both the time and space complexity grow quadratically with sequence length, making it difficult to support the low-latency and high-concurrency requirements of immersive interactive scenarios. At the same time, speech signals, as highly structured sequential data, contain substantial acoustic redundancy. Conventional global attention mechanisms fail to differentiate the actual contribution of individual speech frames to recognition outcomes, leading to unnecessary computational overhead. Inspired by the Informer model in long-sequence prediction, a probabilistic sparse self-attention mechanism was introduced in this model. By measuring the Kullback-Leibler (K-L) divergence between the query-key attention distribution and a uniform distribution, the mechanism identifies and retains only those contextual dependencies that are critical to the speech recognition task. In this way, computational load is significantly reduced, and the system's real-time responsiveness and scalability in VR environments are enhanced. The principle of the attention mechanism is illustrated in Figure 2.

The core of probabilistic sparse self-attention lies in filtering queries with high attention contributions, while performing refined modeling only on key frames. Specifically, for each query vector w_u, the K-L divergence between its attention

distribution over all keys $o(j_k \| w_u)$ and the uniform distribution w is computed. A larger divergence value indicates that the query plays a more important role in self-attention, while smaller values can be regarded as redundant. Concretely, let j_k represent the k-th row of the key matrix, M denote the sequence length, and f represent the input feature dimension. The self-attention of w_u with respect to K can be expressed as:

$$o\left(j_k \| w_u\right) = \frac{r^{w_u j_k^S / \sqrt{f}}}{\sum_{m=1}^{M} r^{w_u j_k^S / \sqrt{f}}} \tag{1}$$

$$ATT(w_u, K, V) = \sum_{k=1}^{M} o\left(j_k | w_u\right) n_k \tag{2}$$

The K-L divergence is defined as:

$$KL\left(w \| o\right) = \ln \sum_{k=1}^{M} e^{\frac{w_u j_k^t}{\sqrt{f}}} - \frac{1}{M} \sum_{k=1}^{M} \frac{w_u j_k^t}{\sqrt{f}} - \ln M \tag{3}$$

The sparsity of w_u can then be computed as:

$$L_{SP}(w_u, K) = \ln \sum_{k=1}^{M} e^{\frac{w_u j_k^T}{\sqrt{f}}} - \frac{1}{M} \sum_{k=1}^{M} \frac{w_u j_k^T}{\sqrt{f}} \tag{4}$$

To further improve the efficiency of query selection, a sampling-based approximation method was employed for the approximate computation of divergence values. Only the top-M_{SP} queries with the highest divergence were retained, while the remaining queries directly output their value vectors without additional attention weighting. Specifically, let the randomly sampled key matrix be denoted as $\tilde{J}$ and the sampled value as $\tilde{M}$. The approximate calculation is expressed as:

$$\bar{L}_{SP}(w_u, \tilde{K}) = \underset{k}{MAX}\left\{\frac{w_u j_k^T}{\sqrt{f}}\right\} - \frac{1}{\tilde{M}} \sum_{j=1}^{\tilde{M}} \frac{w_u j_k^T}{\sqrt{f}} \tag{5}$$

The sampling value $\tilde{M}$ is computed as $\tilde{M} = e_{SA} lnM$, where e_{SA} represents the sampling rate. If u denotes the index of a query within the top-M_{SP} set, the probabilistic sparse self-attention can be formulated as:

$$ATT(w_u, K, V) = \begin{cases} \sum_{k=1}^{M} o\left(j_k | w_u\right) n_k, if \ \ u \in Index_{SP} \\ n_u, else \end{cases} \tag{6}$$

This mechanism preserves sensitivity to critical phonemes, prosodic features, and semantic boundary frames in English speech, while reducing the computational complexity of attention from $O(M^2)$ to $O(M \ log \ M)$. As a result, efficient and smooth end-to-end speech recognition is achieved in VR-based learning environments, providing a reliable and efficient acoustic representation foundation for subsequent pronunciation quality assessment and interaction behavior analysis.

In VR-based English learning, emphasis is placed on real-time human-computer interaction and continuous pronunciation evaluation. Although traditional feed-forward networks possess strong representational capacity, their linear transformation layers contain a large number of parameters and incur high computational costs, making it difficult to achieve low-latency responses on resource-constrained VR devices. Drawing on the successful application of low-rank matrix factorization in speech recognition model compression, the original structure was replaced with a low-rank feed-forward module. In this design, the original weight matrices Q_1 and Q_2 are decomposed into a cascaded form of two low-rank matrices R and F through the introduction of a bottleneck layer. This approach substantially reduces the parameter size and the number of matrix multiplications while maintaining nearly the same representational capacity, thereby supporting efficient and stable operation of the model in VR environments. The low-rank feed-forward module is grounded in parameter approximation and structural reparameterization through matrix decomposition. Specifically, the fully connected layers in the original feed-forward module are replaced with two low-dimensional linear transformations, with a nonlinear activation function embedded between them to form a bottleneck structure in an "expansion-compression" configuration.

Formally, let the weight matrices of the two linear layers be denoted by Q_1 and Q_2, the output dimension of the feed-forward module by f, and the hidden dimension by f_g. The output of the feed-forward module is given by:

$$FFN(A) = Dropout(Swish(AQ_1))Q_2 \tag{7}$$

If the dimension of the bottleneck layer is denoted by f_{ym}, the output of the low-rank feed-forward module is expressed as:

$$S\text{-}FFN(A) = Dropout(Swish(AR_1F_1))R_2F_2 \tag{8}$$

This design assumes that the original weight matrix is inherently low-rank and can be approximated by fusing and reducing cross-channel features through one-dimensional convolution operations. Given an original transformation dimension F, the introduction of an intermediate dimension E (much less than F) reduces the parameter scale from $O(F^2)$ to $O(FE)$. As a result, memory consumption and inference time are significantly reduced, while the capacity to model critical acoustic and linguistic patterns in speech is preserved.

In VR-supported English learning scenarios, speech recognition systems are required to meet both streaming and high-accuracy demands. On the one hand, real-time dialogue and pronunciation correction require the model to support low-latency, online decoding. On the other hand, long-term analysis of learners' pronunciation quality and interaction patterns necessitates recognition outputs of sufficient accuracy. While the CTC decoder enables streaming output through frame-level modeling, its conditional independence assumption constrains contextual modeling capacity, thereby limiting recognition accuracy. In contrast, the attention-based decoder leverages autoregression and the attention mechanism to effectively exploit global information, achieving higher accuracy but facing challenges in direct application to streaming tasks. Inspired by joint CTC and recurrent neural network transducer (RNN-T) decoding, a dual-path decoding mechanism combining CTC and attention-based decoding was introduced to fully integrate their respective strengths. This design preserves the streaming capability while

substantially improving recognition accuracy, thereby providing reliable outputs for real-time instructional interaction and in-depth behavioral analysis in VR environments.

Within this dual-path decoding structure, the CTC decoder functions as the first-pass decoding module, performing preliminary streaming recognition on input speech frame sequences. Prefix Beam Search was adopted for CTC decoding, which mitigates the performance degradation associated with probability splitting in conventional Beam Search by merging hypothesis paths with identical prefixes and accumulating their probabilities. The CTC mechanism allows the insertion of blank symbols between output labels to address the mismatch between speech frames and text label lengths, thereby achieving frame-level alignment for continuous English speech. Its output produces an n-best list of candidate hypotheses, which serves as the input to the second-pass decoder. The CTC branch operates in a streaming mode, ensuring real-time responsiveness to learners' speech input and satisfying the low-latency requirements of VR-based interaction.

The second-pass decoding is performed by the attention-based decoder, which is responsible for rescoring the n-best candidate hypotheses generated by the CTC branch. This decoder employs cross-attention between the high-level acoustic representations produced by the encoder and the text sequences, thereby fully integrating global speech context. Functioning in a role similar to that of a language model, it effectively corrects pronunciation confusions, grammatical errors, and lexical deviations present in the CTC candidates. During training, the system is jointly optimized through multi-task learning with both CTC and attention-based losses. A label smoothing strategy is applied to the attention loss to enhance generalization capacity and improve convergence stability. The final recognition output is determined by combining the CTC hypothesis scores with the rescoring scores from the attention mechanism, where the highest-scoring hypothesis is selected as the final output. Specifically, a hyperparameter η is introduced to balance the relative importance of the CTC and attention losses. Let a denote the input acoustic features and b represent the corresponding output results; the loss function of the dual-path decoding structure is expressed as:

$$LOSS_{D-TD}(a, b) = \eta LOSS_{CTC}(a, b) + (1 - \eta) LOSS_{ATT}(a, b) \tag{9}$$

For the loss function of the attention-based decoder, the objective is to maximize the probability of autoregression. Let i denote the label index of the output; the corresponding expression is formulated as:

$$O\left(b \mid a\right) = \prod_{i} O\left(b_i \mid a, b_{1:i-1}\right) \tag{10}$$

This dual-path decoding mechanism thus combines the streaming capability with the high accuracy. It is particularly well suited to continuous and interactive speech input in VR-based English learning environments. On the one hand, it enables real-time pronunciation feedback and conversational interaction, thereby enhancing immersion and learner engagement. On the other hand, it provides high-quality text output for subsequent pronunciation assessment, error pattern analysis, and interaction behavior mining. In this way, the system not only functions as an effective tool for real-time instruction but also establishes a reliable data foundation for educational researchers to analyze learning behaviors.

2.2 Overall design of the VR-supported immersive English learning system

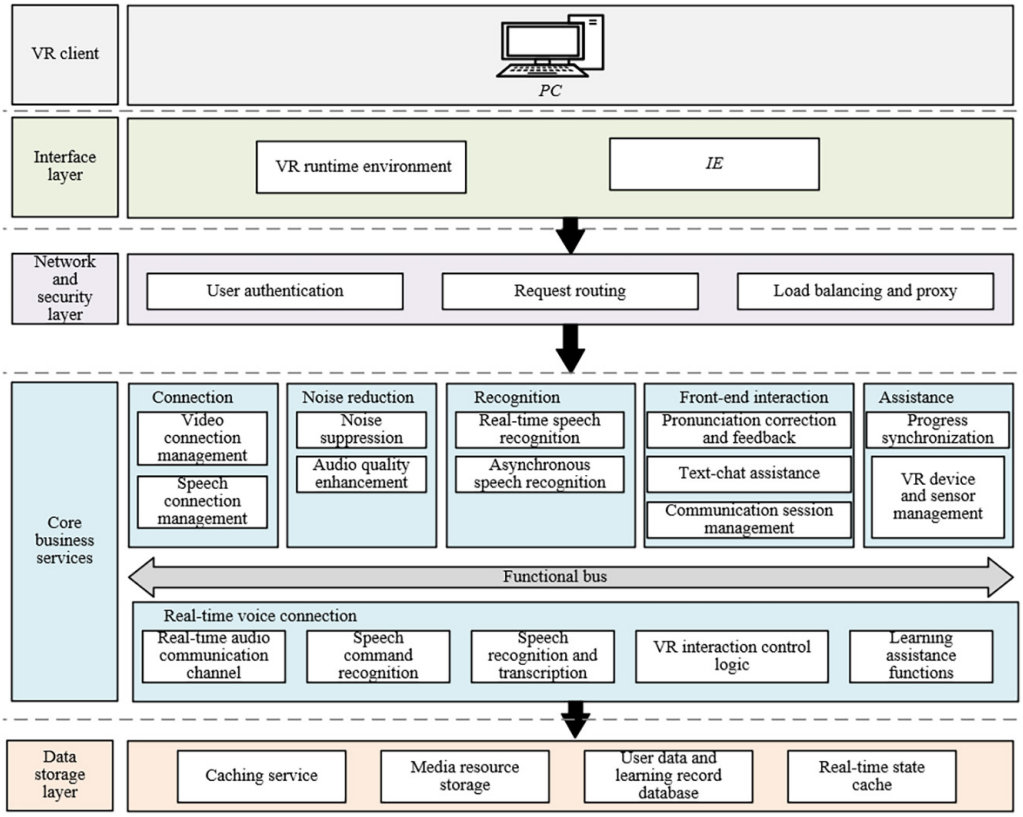

Fig. 3. Software architecture of the VR-supported immersive English learning system

The constructed VR-supported immersive English learning system was designed with a layered architecture to comprehensively support multimodal interaction and intelligent services, as illustrated in Figure 3. The VR client functions as the entry point for user interaction. Through the interface layer, connections are established with the VR runtime environment and inter-process communication mechanisms, which are then integrated with the network and security layer. At this level, fundamental services such as identity authentication, request routing, and load balancing are provided to ensure secure communication and system scalability. Within the core business service layer, modules for speech connection management, real-time noise suppression, and audio quality enhancement are incorporated. These modules not only guarantee low-latency and high-concurrency connections for audio-video communication but also significantly improve the quality of speech input in complex environments through deep learning-based noise suppression and speech enhancement techniques, thereby laying a foundation for subsequent processing.

The recognition module serves as the intelligent core of the system. It integrates the end-to-end speech recognition model and supports both real-time streaming recognition and asynchronous high-accuracy transcription. This design enables immediate pronunciation feedback and speech-interaction responses, while also facilitating offline analysis and evaluation of recorded speech. The front-end interaction module further combines recognition results with educational semantic interpretation, enabling real-time pronunciation error correction, text-chat assistance, and multimodal session management, thereby enhancing the authenticity and effectiveness of instructional interaction. The auxiliary module ensures continuity of the learning state and adaptive interaction through progress synchronization

and VR device-sensor management. All functionalities are encapsulated within the functional bus layer, which integrates real-time audio communication, speech command recognition, speech transcription, VR interaction control, and multiple learning assistance features. These functions are supported by the data storage layer, which provides caching, media resource storage, user and learning record databases, and real-time state caching to achieve efficient data management and persistence. Through the deep integration of speech recognition and VR-based interaction technologies, the system establishes an intelligent English learning environment characterized by low latency, high immersion, and strong feedback capacity.

3 EXPERIMENTS AND ANALYSIS

To evaluate the overall performance of speech recognition models in VR-based English learning environments, several lightweight architectures suitable for embedded deployment were selected for comparison. The experimental results presented in Table 1 demonstrate that the proposed end-to-end speech recognition model achieved the best performance across three critical dimensions: CER (4.52%), RTF (0.078), and memory usage (268 MB). Notably, superior recognition accuracy was obtained while maintaining low computational overhead. Although EfficientConformer approached the proposed model in accuracy, its real-time performance was inferior. ContextNet and QuartzNet demonstrated moderate efficiency but were limited in accuracy, while Speech-Transformer exhibited the highest computational complexity, resulting in poor real-time performance. These findings indicate that the proposed model achieved the most favorable balance between recognition accuracy, response speed, and resource consumption, fully satisfying the stringent requirements of real-time speech interaction in VR-based English learning environments.

Table 1. Performance comparison of end-to-end speech recognition models for VR-based English learning

Model Structure	Character Error Rate (CER) (%)	Real-Time Factor (RTF)	Model Size (MB)	Memory Usage (MB)
QuartzNet-15x5	6.12	0.152	19.2	285
ContextNet	5.83	0.138	22.7	312
EfficientConformer	5.24	0.121	25.3	336
Speech-Transformer	6.05	0.194	31.8	398
Proposed model	4.52	0.078	23.6	268

Table 2. Ablation study of different optimization strategies in the proposed model

Model Configuration	CER (%)	RTF	Parameters (M)	Inference Speed (Frames/ms)
Baseline Conformer	5.18	0.148	29.7	68.2
+ Probabilistic sparse attention	4.87	0.103	27.4	92.5
+ Low-rank feed-forward	5.02	0.121	24.6	79.3
+ Dynamic convolution	4.93	0.116	26.8	83.7
All combined	4.52	0.078	23.6	125.4

To evaluate the contribution of each optimization module within the proposed model, a series of ablation experiments was conducted. The results presented in

Table 2 indicate that the probabilistic sparse attention mechanism contributed most significantly to inference efficiency, reducing RTF from 0.148 to 0.103 while lowering CER by 0.31%. The low-rank feed-forward module primarily reduced the parameter count and memory usage, though it introduced a slight trade-off in recognition accuracy. The dynamic convolution module maintained recognition accuracy while further improving processing speed. When all optimization strategies were applied jointly, the model achieved optimal performance across all metrics. In particular, inference speed was increased by approximately 84%, and the parameter count was reduced by 20.5%. These findings confirm that the combined optimization strategy effectively addresses the trade-off between limited computational resources and high recognition accuracy in VR-based environments, thereby providing an ideal technical foundation for immersive English learning.

To address the dual requirements of streaming recognition and accuracy in VR-based English learning environments, multiple advanced decoding strategies were compared, as shown in Table 3. The results demonstrate that although CTC Greedy Search yielded the lowest latency, its recognition accuracy was relatively poor. The attention decoder achieved higher accuracy but lacked support for streaming recognition. RNN-T provided a more balanced performance in streaming recognition but left room for improvement in accuracy. The CTC/attention rescoring strategy achieved a reasonable trade-off between accuracy and latency. By contrast, the proposed dynamic hybrid decoding strategy integrated the advantages of multiple decoding methods. It achieved the lowest latency of 118 ms while also obtaining the best recognition accuracy of 4.52% CER, fully meeting the needs of real-time speech interaction in VR environments. These results confirm that the dynamic hybrid decoding strategy effectively balances real-time performance with recognition accuracy, making it a reliable decoding solution for immersive English learning.

Table 3. Performance comparison of different decoding strategies in VR-based English learning scenarios

Decoding Strategy	CER (%)	First-Word Latency (ms)	RTF	Streaming Support
CTC Greedy Search	5.46	128	0.062	Yes
Attention decoder	5.12	245	0.196	No
RNN-T	4.98	156	0.105	Yes
CTC/attention rescoring	4.65	132	0.088	Yes
Proposed dynamic hybrid decoding	4.52	118	0.078	Yes

Table 4. Mapping between events and permissions in the immersive VR English learning system

Permission Name	Permission Description	Event	Applicable Role(s)
Virtual tool control	Allows operation of 3D teaching aids, whiteboards, and instructional tools	*VIRTUAL_TOOL_CTRL*	Teacher
Content management	Allows uploading, switching, and deleting learning content	*CONTENT_MANAGE*	Teacher
Speech control	Allows enabling/disabling speech input for specific learners	*VOICE_CTRL*	Teacher and learner
Temporary leave	Removes a user from the current session (reconnection permitted)	*TEMP_LEAVE*	Teacher
Camera control	Allows enabling/disabling user cameras	*CAMERA_CTRL*	Teacher and learner
Model speaker	Grants a learner the permission to perform pronunciation demonstrations	*MODEL_SPEAKER*	Teacher
Group discussion control	Creates or dissolves group discussion sessions	*GROUP_SESSION_CTRL*	Teacher

To clarify the real-time control capabilities required by different instructional roles in VR-based English learning, a WebSocket-based mapping mechanism between permissions and events was designed. As shown in Table 4, virtual tool control and content management constitute the core instructional control functions for teachers, ensuring flexible management of teaching resources. Speech control and model speaker were specifically designed for pronunciation training and classroom interaction, enabling teachers to guide oral activities and designate demonstration speakers, thereby enhancing immersion and engagement. Group discussion control facilitates the organization of collaborative session tasks, aligning closely with the practical needs of group exercises in language learning.

To enable natural and efficient instructional interaction in VR-based English learning environments, a detailed event-driven communication mechanism was designed for virtual teaching tools. The interaction events defined in the immersive VR English learning environment are presented in Table 5. Functionalities such as the 3D handwriting pen, annotation boxes, and voice notes support multimodal content input and annotation, making them particularly suitable for vocabulary explanation, grammar clarification, and pronunciation demonstration. Functions such as highlighting and 3D teaching tool manipulation strengthen instructional guidance, assisting learners in focusing on key information. Features such as undo/redo and stroke saving ensure process traceability and resource reusability, thereby supporting learner review after class as well as teacher reflection.

Table 5. Interaction event definitions for teaching tools in immersive VR English learning environments

Function Name	Function Description	Event	Applicable Role(s)	Supported Scenario(s)
3D handwriting pen	Enables freehand writing in the virtual scene	*VR_PEN_DRAW*	Teacher and learner	Whiteboard and 3D scene
Add an annotation box	Adds text annotations to an object or region	*ADD_ANNOTATION*	Teacher	Any virtual object
Voice note	Attaches voice explanations to specific content	*ADD_VOICE_NOTE*	Teacher and learner	Tools, whiteboard, and exercises
Highlight object	Highlights designated instructional tools or regions	*HIGHLIGHT_OBJECT*	Teacher	3D tools and scene
Undo/redo action	Reverts or restores a previous operation	*UNDO_REDO_ACTION*	Teacher and learner	All editing operations
Clear content	Clears strokes or annotations in the current scene	*CLEAR_CONTENT*	Teacher	Whiteboard and 3D scene
Switch scene	Switches between learning scenes or pages	*SWITCH_SCENE*	Teacher	Entire environment
Save drawing	Saves current strokes as learning resources	*SAVE_DRAWING*	Teacher and learner	Whiteboard and 3D scene
Manipulate 3D tool	Moves, rotates, or scales virtual teaching tools	*MANIPULATE_TOOL*	Teacher and learner	3D tools

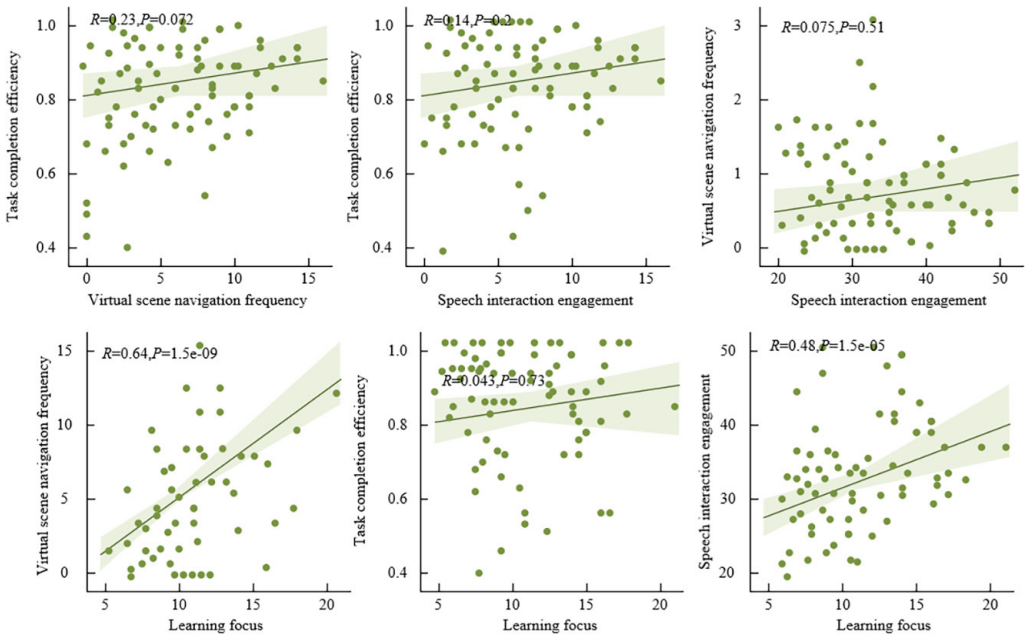

Fig. 4. Correlation analysis results of multimodal interaction behaviors in VR-based English learning environments

To investigate the key behavioral factors influencing learning effectiveness in VR-based English instruction, correlation analyses were conducted across four core interaction dimensions, as shown in Figure 4. The results revealed a strong positive correlation between learning focus and task completion efficiency (R = 0.64, $P = 1.5 \times 10^{-9}$), indicating that sustained attention in VR environments constitutes the most critical factor for improving both the quality and efficiency of task execution. A significant moderate positive correlation was also observed between speech interaction engagement and task completion efficiency ($R = 0.48$, $P = 1.5 \times 10^{-5}$), confirming that active oral practice contributes substantially to learning outcomes. By contrast, virtual scene navigation frequency demonstrated only a weak and statistically insignificant correlation with learning focus (R = 0.075, P = 0.51), suggesting that exploratory behaviors do not necessarily translate into effective learning engagement. Excessive or unguided navigation may even divert attention. Notably, although the correlation coefficient between speech interaction engagement and learning focus reached R = 0.45, the associated P value (0.93) was far above the threshold for statistical significance. This finding suggests that a stable linear relationship may not exist between the two variables, or that a larger dataset is required for reliable validation. Taken together, these findings highlight that the design of immersive VR English learning environments should prioritize narrative and task structures that sustain and enhance learner focus, while strongly encouraging speech production. Navigation functions should be guided by explicit learning objectives so that exploration becomes purposeful and beneficial, rather than a potential source of distraction.

4 CONCLUSION

This study was conducted around the theme of designing a VR-supported immersive English language learning environment and analyzing interaction behaviors. An integrated VR-based system combining end-to-end speech recognition with

multimodal interaction was constructed, and learner behavior patterns and learning outcomes were systematically examined. In terms of research content, a lightweight speech recognition model incorporating a low-rank feed-forward module, probabilistic sparse attention, and a dual-path CTC/attention decoding mechanism was innovatively introduced. This design effectively addressed the challenges of real-time processing and recognition accuracy for long speech sequences in VR environments. In addition, the overall architecture of the immersive learning system was designed and implemented, enabling core functions such as virtual tool manipulation, real-time speech interaction, and scenario-based learning tasks. In terms of research findings, experimental validation demonstrated that the proposed model achieved superior performance compared to baseline models in recognition accuracy, real-time responsiveness, and resource efficiency. Furthermore, analysis of behavioral data identified three typical learner groups—high-efficiency, moderate, and struggling—and revealed that learning focus and speech interaction engagement were the key behavioral factors influencing learning effectiveness. These findings indicate that the study provides both an effective technical solution and theoretical insights for achieving efficient and adaptive immersive language learning.

The principal contribution of this study lies in the first deep integration of a lightweight speech recognition model with a VR learning environment, coupled with a data-driven approach that revealed intrinsic connections between learner behaviors and learning outcomes. This integration provides an important paradigm for linking academic research with practical applications in the field of intelligent education. Nevertheless, several limitations remain. First, the relatively limited sample size may constrain the generalizability of the behavior classification model. Second, discrepancies between the controlled experimental setting and the complexity of real teaching environments may affect the validation of long-term learning outcomes. Third, the current system has not yet fully realized adaptive regulation with respect to individual learner differences. Future work will focus on several directions. Expanding the scale and diversity of experimental samples will be prioritized to enhance the universality and robustness of the proposed model. Reinforcement learning mechanisms will be explored to enable real-time perception of learner states and adaptive content recommendation. The scope of multimodal data analysis will be broadened by incorporating eye-tracking, gesture, and other behavioral signals, thereby providing a more comprehensive assessment of learning engagement and cognitive load. Finally, the potential of this system will be investigated in more diverse language learning contexts, with the aim of advancing immersive language learning toward greater personalization and intelligence.

5 REFERENCES

[1] R. Alsalameen, L. Almazaydeh, B. Alqudah, and K. Elleithy, "Information technology students' perceptions toward using virtual reality technology for educational purposes," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 7, pp. 148–166, 2023. https://doi.org/10.3991/ijim.v17i07.37211

[2] D. A. Wiliyanto, Gunarhadi, F. K. Anggarani, J. Yuwono, and A. Anggrellangi, "Design of DSLIs based on virtual reality for deaf students," *Ingénierie des Systèmes d'Information*, vol. 30, no. 1, pp. 267–278, 2025. https://doi.org/10.18280/isi.300123

[3] H. D. Sharma, Y. Misra, S. Kumar, B. M. Rao, and B. Ch, "Expanding an education-based collision detection system created on virtual reality and augmented reality," *International Journal of Interactive Mobile Technologies*, vol. 17, no. 17, pp. 108–120, 2023. https://doi.org/10.3991/ijim.v17i17.42831

[4] R. Speidel, E. Felder, A. Schneider, and W. Öchsner, "Virtual reality against Zoom fatigue? A field study on the teaching and learning experience in interactive video and VR conferencing," *GMS Journal for Medical Education*, vol. 40, no. 2, p. Doc19, 2023. https://doi.org/10.3205/zma001601

[5] K. Abramczuk, Z. Bohdanowicz, B. Muczyński, K. H. Skorupska, and D. Cnotkowski, "Meet me in VR! Can VR space help remote teams connect: A seven-week study with Horizon Workrooms," *International Journal of Human-Computer Studies*, vol. 179, p. 103104, 2023. https://doi.org/10.1016/j.ijhcs.2023.103104

[6] C. Moreira *et al.*, "Toward VR in VR: Assessing engagement and social interaction in a virtual conference," *IEEE Access*, vol. 11, pp. 1906–1922, 2022. https://doi.org/10.1109/ACCESS.2022.3233312

[7] S. Somadayo and M. Jamil, "Evaluation of online learning system based on virtual reality technology for developing reading literacy in elementary schools," *Mathematical Modelling of Engineering Problems*, vol. 12, no. 5, pp. 1731–1740, 2025. https://doi.org/10.18280/mmep.120528

[8] D. Novaliendry, W. Z. Syaputra, A. D. Samala, and R. Marta, "Design of a virtual simulation for Tsunami disaster education and mitigation at Teluk Penyu Beach, Cilacap," *Journal Européen des Systèmes Automatisés*, vol. 58, no. 6, pp. 1257–1263, 2025. https://doi.org/10.18280/jesa.580615

[9] R. Safdari, M. G. Saeedi, A. Valinejadi, H. Bouraghi, and H. Shahnavazi, "Factors affecting on deployment of speech recognition technology in health care organizations of Iran," *International Journal of Computer Science and Network Security*, vol. 17, no. 1, pp. 35–41, 2017.

[10] H. Hermansky, J. R. Cohen, and R. M. Stern, "Perceptual properties of current speech recognition technology," *Proceedings of the IEEE*, vol. 101, no. 9, pp. 1968–1985, 2013. https://doi.org/10.1109/JPROC.2013.2252316

[11] H. Eftekhari, "Transcribing in the digital age: Qualitative research practice utilizing intelligent speech recognition technology," *European Journal of Cardiovascular Nursing*, vol. 23, no. 5, pp. 553–560, 2024. https://doi.org/10.1093/eurjcn/zvae013

[12] X. Chai, J. Yang, and Y. Liu, "Influential factors for medical students' classroom concentration—evaluation with speech recognition and face recognition technology," *BMC Medical Education*, vol. 24, no. 1, p. 1236, 2024. https://doi.org/10.1186/s12909-024-06204-5

[13] H. D. J. Jeong, S. K. Ye, J. Lim, I. You, and W. Hyun, "A computer remote control system based on speech recognition technologies of mobile devices and wireless communication technologies," *Computer Science and Information Systems*, vol. 11, no. 3, pp. 1001–1016, 2014. https://doi.org/10.2298/CSIS130915061J

[14] O. Z. Mamyrbayev, K. Alimhan, B. Amirgaliyev, B. Zhumazhanov, D. Mussayeva, and F. Gusmanova, "Multimodal systems for speech recognition," *International Journal of Mobile Communications*, vol. 18, no. 3, pp. 314–326, 2020. https://doi.org/10.1504/IJMC.2020.107097

[15] S. Liu *et al.*, "Recent progress in the CUHK dysarthric speech recognition system," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2267–2281, 2021. https://doi.org/10.1109/TASLP.2021.3091805

[16] F. H. Khoso, D. N. Hakro, and S. Z. Nasir, "Sindhi speech recognition system," *International Journal of Computer Science and Network Security*, vol. 19, no. 11, pp. 21–28, 2019.

[17] F. Barkani, M. Hamidi, O. Zealouk, and H. Satori, "Speech recognition algorithms-based cough recognition system," *International Journal of Online and Biomedical Engineering*, vol. 19, no. 12, pp. 49–61, 2023. https://doi.org/10.3991/ijoe.v19i12.40471

6 AUTHORS

Lanlan Fang, is an Associate Professor. She graduated from Xiangtan University in 2004, majoring in English. And she received a master's degree from Hunan Normal University in 2015. She is currently an English teacher at the Public Course Department of Hunan Mechanical & Electrical Polytechnic. Her research interests include English language teaching, information-based teaching, and vocational education internationalization as well. She has published more than twenty academic papers, published two academic monographs, led a provincial-level educational research project and two prefecture-level educational research projects. She has won two national awards and more than fifteen provincial awards in the teaching competitions (E-mail: fanglanlan1982@163.com).

Xiujuan Wang is an Associate Professor. She received the Master's degree from Shanghai Maritime University, P.R. China. She is currently a Lecturer in Hunan Railway Professional Technology College, China. The main courses she teaches include college English, Computer English, Business English, and General Chinese culture. Her research interests include English teaching, information-based teaching and second language acquisition and international studies as well (E-mail: p121690@siswa.ukm.edu.my).

Liang Zhang received the Master's degree from Central South University, P.R. China. She is currently an Associate Professor at the Hunan Mechanical & Electrical Polytechnic, China. Her research interests include vocational English teaching and vocational education internationalization as well (E-mail: lenazhang103@126.com).