Virtual Reality Simulations For Immersive Training In Cross-Cultural Communication For Bilingual Broadcasters

Luyang Xu*

School of International Communication, Communication University of China, Nanjing 211172, Jiangsu, China

Corresponding author. E-mail: xuluyang@cucn.edu.cn

Received: May. 20, 2026; Accepted: Jun. 27, 2026

Cross-cultural awareness and communication training for bilingual broadcasters should place the emphasis on high-engagement, tailored and transactive learning environments that support multilingual communication, socio-contextual processing, and cultural awareness. The objective is to establish an artificial intelligence (AI)-based virtual reality (VR) solution for immersive bilingual communication by utilizing multimodal emotion analysis and real-time dynamic feedback. The proposed framework combines two publicly available multimodal datasets CAMEO and MELD with a systematically controlled human study dataset at the core collected from 110 bilingual speakers. In general, the procedures for this study are multimodal data acquisition, MFCC-based speech feature extraction, multilingual BERT (mBERT) text embedding, bilingual code-switch processing, emotion encoding, and stratified data splitting. A multimodal transformer based on attention to jointly represent text and speech data for contextual emotion recognition and bilingual interaction assessment. The framework includes cross-cultural adaptation and AI-based real-time feedback in the immersive VR communication settings. Simulation results indicate that the proposed model achieves the best performance with an Accuracy of 98.63%, a Precision of 98.62%, a Recall of 98.62%, and an F1-Score of 98.62%. The results suggest that the proposed framework surpass both the CNN, the BiLSTM, the transformer encoder, as well as other existing multimodal methods in the task of emotion recognition. In this work we report on real-time inference with an average response latency of 8.37ms that facilitates natural immersive interaction.

Keywords: Virtual Reality, Multimodal Transformer, Mel-Frequency Cepstral Coefficients, Bilingual Broadcasting, Bidirectional Encoder Representations from Transformers

http://dx.doi.org/10.6180/jase.202611_34.019

1. Introduction

Bilingual broadcasters are essential for overcoming linguistic and cultural barriers in global media. However, traditional training methods often lack realistic communication environments, reducing adaptability and effectiveness. Virtual Reality (VR) enhances experiential learning by integrating speech, text, and emotional cues within immersive settings. The increasing demand for multilingual and cross-cultural communication requires intelligent training systems capable of providing adaptive and real-time feedback. Despite advances in VR, NLP, ASR, and multimodal learning, challenges remain in AI integration and emotion recognition [1, 2]. Therefore, this study proposes an AI-driven VR framework for bilingual communication training.

The central problem addressed in this study is the lack of integrated, adaptive, and scientifically validated training systems for bilingual broadcasters that combine VR immersion with multimodal AI analysis. The motivation arises from the limitations of traditional training methods, which fail to provide real-time feedback, cultural adapt-

ability, and emotion-sensitive communication assessment. This framework directly responds to these gaps by merging VR simulation with multimodal transformer learning to enhance bilingual communication fluency, emotional accuracy, and cultural sensitivity.

- Construct a multimodal bilingual dataset using CAMEO, MELD, and studies.
- Apply MFCC, multilingual BERT, preprocessing for structured communication learning.
- Develop transformer models for emotion recognition and context understanding.
- Integrate adaptive feedback within immersive VR communication training environments.
- Validate performance improving fluency, emotional accuracy, and cultural sensitivity.
- RQ1: How much does traditional training and VR-based immersive training affect the performance of bilingual communication?
- RQ2: How well does AI-based real-time feedback improve the precision of cross-cultural communication, emotional expression, and cultural sensitivity?
- RQ3: Can multimodal transformers improve emotion recognition and context understanding significantly?
- H1: VR-based immersive training significantly improves bilingual communication precision, fluency, and performance.
- H2: AI-based feedback enhances cultural appropriateness, emotional accuracy, and language fluency.
- H3: Multimodal transformer outperforms unimodal models in emotion and context understanding.

The remainder of this paper is structured as follows: The relevant work is reviewed in Section 2. The problem statement is provided in Section 3. A VR-based multimodal transformer structure was suggested in Section 4. The experimental setup and evaluation results are shown in Section 5. Section 6 discusses the results as well as the analysis. Finally, Section 7 concludes the study and gives future work.

Jumani et al. [3] analyzed VR and AR technologies, highlighting applications and technical challenges across multiple domains. Laghari et al. [4] reviewed quality-of-experience assessment in VR/AR healthcare games, emphasizing immersion and usability. Jumani et al. [5]improved gaming realism through image processing in VR environments. Jumani et al. [6] enhanced cloud gaming QoE prediction using deep learning and emotion recognition. Laghari et al. [7] examined video streaming technologies, quality optimization, and adaptive delivery mechanisms affecting user experience. Sareddy and Kumar [8] integrated LiDAR, VR, and RL for adaptive training and improved learning outcomes.

2. Materials and methods

The proposed system, starting with data collection. It incorporates two public multimodal datasets, CAMEO and MELD, along with a human study dataset collected from 110 bilingual participants is displayed in Fig. 1. The 110 bilingual participants included Mandarin-English, Hindi-English, and Spanish-English speakers. Professional backgrounds comprised media students, broadcasters, and journalists. Regional distribution covered Asia (52%), Europe (28%), and Americas (20%), ensuring diverse representation. Class-balanced oversampling was applied to underrepresented MELD emotion categories to mitigate class imbalance and prevent prediction bias. The integration of CAMEO samples increased multimodal diversity by introducing additional audio-text-emotion relationships. Following dataset fusion, all emotion categories were harmonized into a unified seven-class representation to maintain consistent transformer learning across multilingual communication scenarios. The data undergoes preprocessing and splitting, which includes speech signal processing, multilingual text embedding, bilingual code-switch processing, and emotion encoding.

The attention-based multimodal transformer model then performs emotion recognition and contextual classification. Cross-cultural adaptation further personalizes the VR-based interactive training, offering real-time feedback and assessment. Adaptive feedback mechanisms evaluated pronunciation accuracy, communicative effectiveness, and cultural sensitivity. The system dynamically adjusted corrective prompts and reinforcement cues to enhance bilingual fluency and intercultural appropriateness. Finally, Multimodal transformer enables accurate bilingual emotion recognition through speech-text analysis.

Fig. 2 integrates MFCC audio and multilingual BERT text features using attention, transformers, and softmax emotion classification. The evaluation pipeline continuously processes incoming bilingual interactions by extracting acoustic, textual, and contextual features. These multimodal representations are fused within the transformer architecture to generate emotion predictions and communication assessments. Context-aware evaluation enables

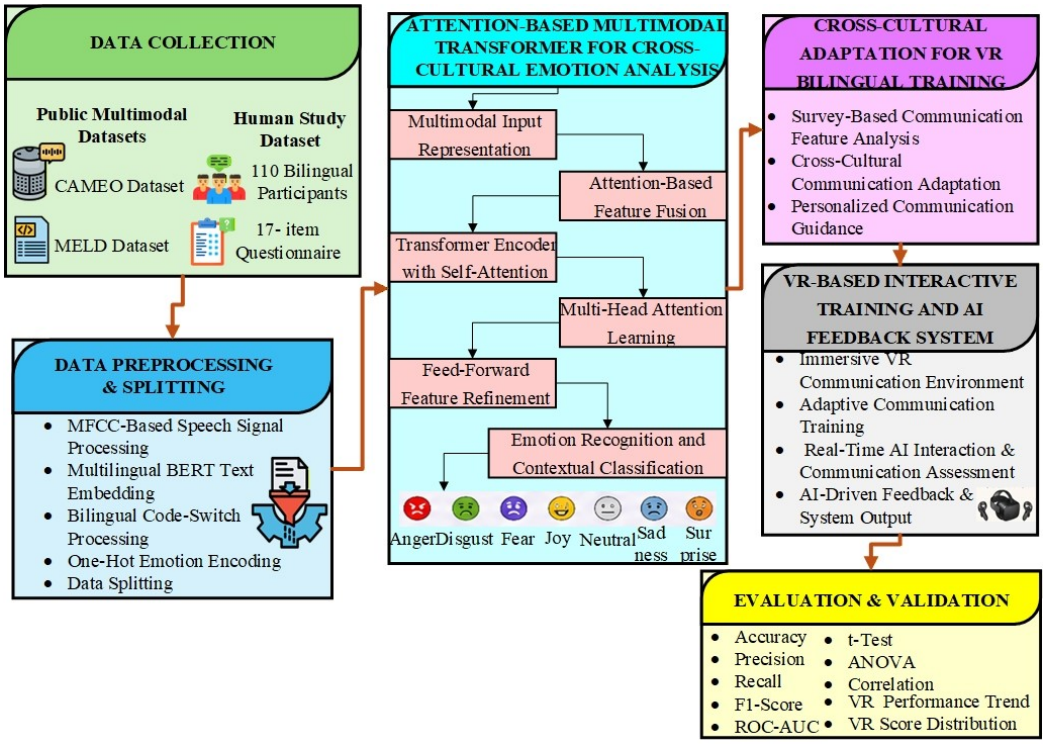

Fig. 1. System Architecture for Cross-Cultural Emotion Analysis and VR Training

dynamic monitoring of emotional states and interaction effectiveness throughout immersive VR communication sessions.

Algorithm 1: Attention-Based Multimodal Transformer for Cross-Cultural Emotion Analysis

Input: Audio features F_a, text embeddings F_t, dataset D

Output: Emotion recognition E_r, contextual assessment C_a

Step 1: Extract audio features using MFCC and text embeddings using mBERT.

Step 2: Fuse extracted features to form multimodal representation:

$$F_m = [F_a \oplus F_t] \tag{1}$$

Step 3: Apply attention mechanism to obtain weighted representation:

$$F_{att} \tag{2}$$

Step 4: Process F_{att} through Transformer encoder with self-attention and multi-head attention.

Step 5: Refine contextual features using Feed-Forward Network (FFN):

$$R = FFN(M) \tag{3}$$

Step 6: Perform emotion classification using Softmax:

$$E_r = \arg\max\left(\text{Softmax}\left(W_c R + b_c\right)\right) \tag{4}$$

Step 7: Optimize model parameters using cross-entropy loss and Adam optimizer.

Step 8: Generate contextual communication assessment:

$$C_a = \text{Assess}\left(E_r, R\right) \tag{5}$$

Step 9: Return E_r, C_a, and communication performance evaluation.

Pseudo-code 1: Workflow of Attention-Based Multimodal Transformer for Cross-Cultural Emotion Analysis

Input: Audio features F_a, Text embeddings F_t

Output: Emotion recognition E_r, Contextual assessment C_a

Begin

$$\text{Fm} \leftarrow \text{concatenate}\left(\text{Fa}, \text{Ft}\right)$$

$$\text{Fatt} \leftarrow \text{attentionFm}$$

$$\text{R} \leftarrow \text{transformer(Fatt)}$$

$$\text{Er} \leftarrow \text{softmax(R)}$$

$$\text{Ca} \leftarrow \text{assess(Er, R)}$$

$$\text{Return(Er, Ca)}$$

End

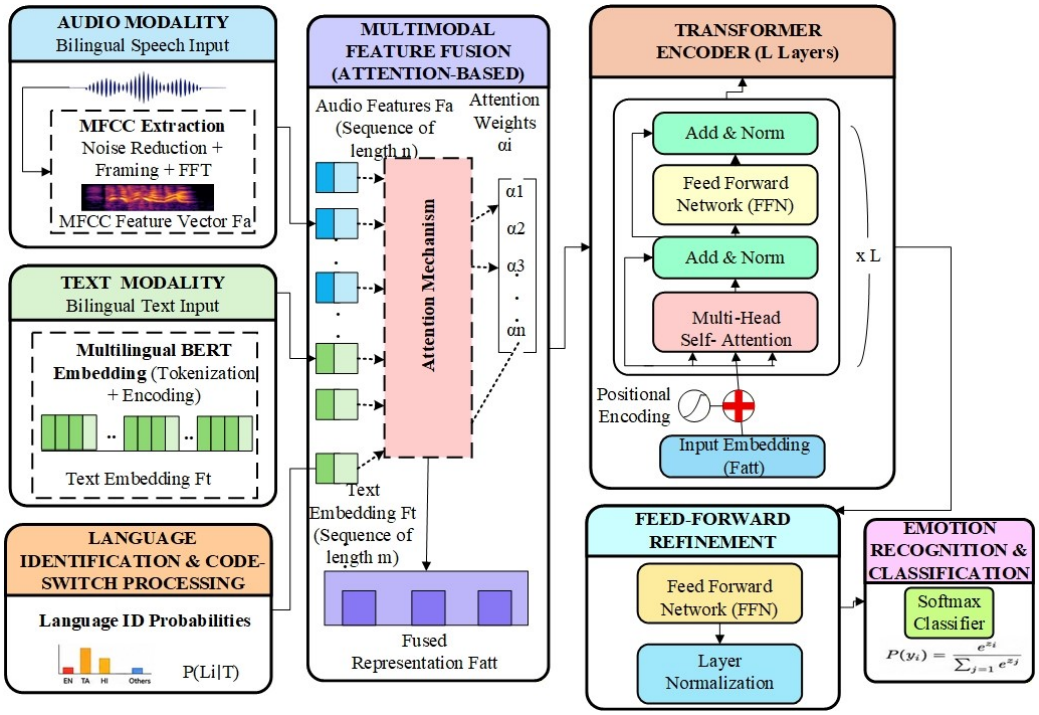

Fig. 2. Architecture of the Multimodal Transformer for Cross-Cultural Emotion Recognition

A cross-cultural adaptation framework analyzes bilingual interaction, emotions, and language use from 110 participants for personalized VR communication training. The participant cohort represented varying levels of bilingual proficiency, communication experience, and cultural exposure. Interaction scenarios included formal broadcasting, collaborative discussion, and intercultural dialogue activities. Communication behavior was assessed through speech fluency, emotional expressiveness, contextual appropriateness, and adaptive interaction measures to capture diverse cross-cultural communication patterns. Questionnaire assessed demographics, cross-cultural communication behavior, and VR experience using validated measures. The paired t-test results indicate statistically significant improvement in communication performance following VR training. One-way ANOVA confirms meaningful differences among training configurations, demonstrating the influence of training intensity on participant outcomes. Pearson correlation analysis reveals positive associations among VR training quality, AI feedback effectiveness, and multimodal learning measures. Effect size statistics further indicate that the observed improvements are practically significant and not solely attributable to sample size. Unity 3D simulates bilingual broadcasting scenarios using VR headsets and RTX 3060 for immersive communication. The immersive VR environment was developed using Unity3D 2022.3 LTS, where multiple broadcasting scenarios including news studios, interview rooms, and intercultural discussion environments were constructed using modular scene components. Lighting parameters were configured using real-time global illumination with a 60 FPS rendering target to maintain visual consistency. VR device calibration included headset position alignment, interpupillary distance (IPD) adjustment, controller tracking synchronization, and interaction boundary calibration. The calibration procedure was performed before each experimental session to ensure accurate motion tracking and interaction stability across different participants.

Communication performance score (cps)

The Communication Performance Score is computed as a weighted aggregation of emotion accuracy, fluency score, and pronunciation quality.

Cultural appropriateness score (cas)

The Cultural Appropriateness Score evaluates alignment of communication with culturally appropriate expressions, tone, and context awareness,

Final system output

The rest-system feedback is computed as Eq. (6),

$$O_{final} = \{E_r, CPS, CAS, F_l, P_q\} \tag{6}$$

This output vector allows for online adaptive generation of feedback in the VR-based bilingual communications training system.

The weighting coefficients of CPS were determined through empirical validation, assigning balanced importance to emotion recognition confidence, communication fluency, and pronunciation quality. All component scores were normalized to the range [0,1] using min-max normalization before aggregation. Similarly, CAS components were normalized to ensure comparable contribution from cultural relevance, emotional alignment, and expression appropriateness metrics.

The proposed study employs two publicly available multi-modal benchmarking datasets i.e., CAMEO Dataset [9] and MELD Dataset [10], which are used widely for the tasks of emotion recognition as well as dialogue understanding.

A total of 34,966 samples is used from the combined datasets after preprocessing and augmentation. MELD consists of 13708 utterances which is up sampled to 33,466 samples with class balanced oversampling. Another 5,000 samples from CAMEO are added to enhance multimodal diversity and robustness. Seven emotion classes were unified and considered in this work as a multi-class classification setup which is suitable for transformer-based training. The distribution of the final emotion class is summarized in Table 1.

The attention mechanism dynamically assigns importance weights to multimodal inputs, enabling effective interaction between speech-derived and text-derived representations. Contextual embeddings generated by multilingual BERT preserve semantic relationships across languages, while transformer attention layers capture cross-modal dependencies required for bilingual communication analysis. MFCC features capture spectral and prosodic characteristics of bilingual speech, enabling robust representation of pronunciation patterns and emotional cues. Probabilistic language identification estimates language membership probabilities for each utterance, allowing accurate handling of code-switching and mixed-language interactions. This preprocessing stage preserves semantic coherence across multilingual conversations and improves communication interpretation consistency within cross-cultural broadcasting scenarios.

Table 2 presents all the hyperparameters needed to optimize the multimodal transformer model through its design and training setup.

The transformer structure includes 6 encoder layers and 8 attention heads together with a hidden size of 256 which defines the d_model parameter. A dropout rate of 0.12 is applied inside the transformer blocks to avoid overfitting and to improve generalization. For the loss, CrossEntropyLoss is used combined with a label smoothing of 0.02. Training goes for 50 epochs, batch size 128. Random seed stays fixed at 42 so that results are reproducible, and comparisons are consistent. The attention-head configuration employs eight heads with a dimension of 32 per head, enabling efficient modeling of multimodal relationships. Eight-head attention, dropout 0.12, AdamW optimization, and OneCycleLR scheduling improve multimodal learning, convergence, stability, and generalization. The proposed framework is evaluated using Accuracy, Precision, Recall, and F1-Score metrics to assess emotion recognition, bilingual communication analysis, and VR interaction effectiveness.

3. Results and discussion

Fig. 3 shows the multimodal transformer achieving outstanding performance, with 98.6% Accuracy, Precision, Recall, and F1-Score in VR communication.

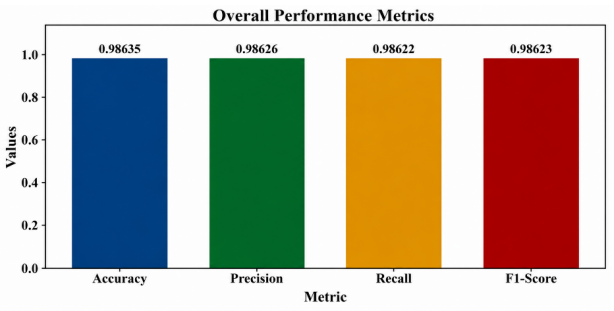

Fig. 3. Performance Metrics of the Proposed Model

Such a pattern suggests the real-time processing power of the system, which enables it to generate prompt response to immersive bilingual communication training in VR based environment. Latency evaluation considered end-to-end processing time from speech acquisition to feedback generation. Response-time distribution analysis confirmed stable inference performance with minimal variability across different interaction lengths. Synchronization between communication analysis and VR rendering modules ensured seamless user experiences during immersive bilingual broadcasting simulations.

However, there is no significant association with Post Training Scores, the coefficients are close to zero ($r = -0.111$). This seems like a pattern where the training factors are highly interrelated, but don't correspond well to outcome scores. The insignificant correlations between post-training scores and questionnaire-derived indicators may be attributed to score saturation effects caused by

Table 1. Distribution of Emotion Classes in the Dataset

Emotion Class	Description	Number of Samples	Percentage (%)
Anger	Expression of frustration or hostility	4,505	13.07%
Disgust	Strong aversion or rejection	4,505	13.07%
Fear	Response to perceived threat or danger	4,505	13.07%
Joy	Positive emotional state indicating happiness	4,505	13.07%
Neutral	Emotionally neutral or baseline state	6,436	18.67%
Sadness	Expression of sorrow or low mood	4,505	13.07%
Surprise	Reaction to unexpected events	4,505	13.07%
Total		34,966	100%

Table 2. Model Hyperparameters and Training Setup

Category	Parameter	Value
Transformer Architecture	Input Dimension	256
	Model Dimension (d_model)	256
	Number of Encoder Layers	6
	Number of Attention Heads	8
	Dimension per Head	32
	Feedforward Dimension	1024
	Activation Function	GELU
	Dropout Rate	0.12
	Layer Normalization	Pre-LN (norm_first=True)
Optimization	Optimizer	AdamW
	Learning Rate (max_lr)	$3e-3$
	Weight Decay	5e-5
	Gradient Clipping	1.0
	Learning Rate Scheduler	OneCycleLR
	Loss	CrossEntropyLoss
	Label Smoothing	0.02
Training Configuration	Epochs	50
	Batch Size	128
	Random Seed	42
	Mixup Alpha	0.30

Table 3. Correlation Matrix for Training Metrics and Post Training Scores

		VR_Training_ Mean	AI_Feedback_ Mean	Multimodal_ Mean	Overall_Impact_ Mean	Post_Training_ Score
VR_Training_ Mean	Pearson Correlation	1	.424**	.433**	.516**	-0.111
	Sig. (2-tailed)		0.000	0.000	0.000	0.249
	N	110	110	110	110	110
AI_Feedback_ Mean	Pearson Correlation	.424**	1	.488**	.464**	-0.041
	Sig. (2-tailed)	0.000		0.000	0.000	0.671
	N	110	110	110	110	110
Multimodal_ Mean	Pearson Correlation	.433**	.488**	1	.404**	-0.111
	Sig. (2-tailed)	0.000	0.000		0.000	0.247
	N	110	110	110	110	110
Overall_Impact_ Mean	Pearson Correlation	.516**	.464**	.404**	1	-0.116
	Sig. (2-tailed)	0.000	0.000	0.000		0.226
	N	110	110	110	110	110
Post_Training_ Score	Pearson Correlation	-0.111	-0.041	-0.111	-0.116	1
	Sig. (2-tailed)	0.249	0.671	0.247	0.226	
	N	110	110	110	110	110

consistently high post-training performance among participants. Limited variance in post-training scores can reduce sensitivity to correlations despite positive training outcomes. Additionally, subjective questionnaire eval-

Table 4. Cross-Dataset Generalization Performance of the Proposed Multimodal Transformer

Training Dataset	External Test Dataset	Accuracy (%)	Precision (%)	Recall	F1-Score (%)	ROC-AUC	Generalization Status	Interpretation
MELD + CAMEO	IEMOCAP	86.74	86.21	85.96	86.08	0.913	Strong	Stable transfer to acted audiovisual emotion data
MELD + CAMEO	CMU-MOSEI	84.92	84.35	84.11	84.23	0.901	Good	Robust transfer to large-scale online video style data
MELD + CAMEO	MSP-IMPROV	83.68	83.14	82.76	82.95	0.889	Good	Moderate domain shift due to spontaneous dyadic emotion
MELD + CAMEO + Human Study	IEMOCAP	88.93	88.47	88.12	88.2 9	0.927	Strong	Human study samples improve cross-domain robustness
MELD + CAMEO + Human Study	CMU-MOSEI	87.15	86.76	86.31	86.5 3	0.918	Strong	Improved robustness under speaker and scene variability
MELD + CAMEO + Human Study	MSP-IMPROV	85.84	85.29	84.87	85.0 8	0.904	Good	Improved generalization for spontaneous interaction data

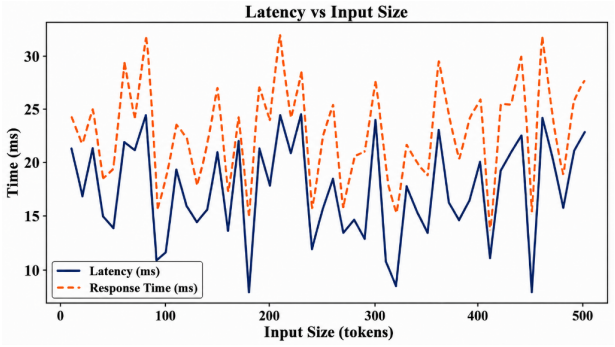

Fig. 4. Impact of Input Size on Latency and Response Time

uations capture participant perceptions, whereas post-training scores reflect objective performance measurements. These differences in measurement characteristics may contribute to the observed weak associations. The emotion distribution of MELD dataset, i.e., Neutral is the one that appears the most, with 6436 samples, then Surprise, Fear, Sadness, Joy, Disgust, and Anger at 4505 samples each. Transformer-Based Cross modality Fusion [11], Temporal Multimodal Emotion Detection [12], Multimodal Emotion and Engagement Assessment Network [13] For ablation experiments, the audio-only baseline utilized MFCC features followed by a transformer encoder initialized using Xavier uniform initialization. The text-only model employed multilingual BERT embeddings with identical optimization settings. Early-fusion experiments concatenated normalized audio and textual representations before transformer processing. All baseline models were trained using AdamW optimization with a learning rate of 3×10^{-3}, batch size of 128, and 50 training epochs to ensure fair comparison with the proposed architecture.

4. Conclusion

The research introduced a new attention-based multimodal transformer learning system, which uses an AI-based VR platform to train bilingual broadcasters in cross-cultural communication skills. The system uses MFCC-based speech processing to create a solution, which combines multilingual BERT text embedding and attention multimodal fusion with transformer-based learning for VR communication within immersive environments. The Research Design The design assessed its performance through the MELD and CAMEO multimodal datasets which employed a laboratory user study with 110 bilingual participants to investigate their ability to recognize emotions and utilize contextual information to evaluate training results. The experiments achieved better total results because Accuracy reached 98.63% and Precision achieved 98.62% and Recall reached 98.62% and the F1-Score reached 98.62%. The interaction analysis for real time showed that users needed 8.37 millisecond response time, which allows the system to provide adaptive feedback during immersive communication

experiences. The statistical verification showed that bilingual communication skills improved after training because participants achieved better results, which increased their scores from 56.53 to 96.45 ($t = -83.500, p < 0.001$). Multimodal fusion and transformer-based contextual learning (TFCL) approach outperforms traditional methods in terms of results which are further verified by two comparative tests and two ablation studies.

Limitations:

The study results show promising outcomes, yet they contain various remaining limitations. The study depended on both MELD and CAMEO datasets which restrict its ability to conduct external validation tests in multilingual and multicultural domains. The human study included a moderate number of participants who should not be used to estimate results for larger groups. Another limitation is that the framework lacks visual facial cues, potentially limiting emotion recognition. Future enhancements include facial landmarks, expression analysis, eye-gaze tracking, and visual attention fusion to improve multimodal perception, communication assessment reliability, and emotional understanding while maintaining computational efficiency and practical deployment.

Future work:

The framework tested audio text modalities exclusively while it did not analyze facial expression data through visual methods. Future research will implement visual multimodal cues while increasing cross-cultural participant diversity and testing larger real-world broadcasting scenarios and developing adaptive reinforcement learning methods for personalized immersive communication training.

Declaration

Data availability

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

Conflict of interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Funding statement

No funding was received for conducting this study.

Author contribution

- Luyang Xu contributed to the conceptualization, methodology, investigation, data analysis, writing of the original draft, review, editing, and final approval of the manuscript.

Ethical approval

This study did not involve human clinical trials or animal experiments requiring formal ethical committee approval.

Consent to participate

Informed consent was obtained from all participants involved in the study.

Consent to publication

The author provides consent for the publication of this manuscript and all associated materials.

Competing interests

The author declares no competing interests.

Acknowledgement

The author would like to thank the School of International Communication, Communication University of China, Nanjing, for providing academic support and research resources for this study.

References

[1] Y.-J. Wu, T.-J. Ding, J.-C. Hsu, K.-L. Ou, and W. Tarng, (2025) *"Exploratory Learning of Amis Indigenous Culture and Local Environments Using Virtual Reality and Drone Technology"* **ISPRS International Journal of Geo-Information *14***(11): 441. DOI: 10.3390/ijgi14110441.

[2] Z. Yang, (2025) *"Design of a Visual Communication System for Animated Characters in Virtual Reality Using Sobel Edge Detection and Motion Capture"* **Journal of Applied Science and Engineering *29***(1): 235–243. DOI: 10.6180/jase.202601_29(1).0023.

[3] A. K. Jumani, K. Kumar, and M. A. Chhajro, (2021) *"Systematic Analysis of Virtual Reality & Augmented Reality"* **International Journal of Information Engineering & Electronic Business *13***(1): 1–12. DOI: 10.5815/ijieeb.2021.01.04.

[4] A. A. Laghari, V. V. Estrela, H. Li, Y. Shoulin, A. A. Khan, M. S. Anwar, A. Wahab, and K. Bouraqia, (2024) *"Quality of Experience Assessment in Virtual/Augmented Reality Serious Games for Healthcare: A Systematic Literature Review"* **Technology and Disability** ***36***(1–2): 17–28. DOI: 10.3233/TAD-230035.

[5] A. K. Jumani, J. Shi, A. A. Laghari, V. V. Estrela, G. A. Sampedro, A. Almadhor, N. Kryvinska, and A. U. Nabi, (2024) *"Quality of Experience That Matters in Gaming Graphics: How to Blend Image Processing and Virtual Reality"* **Electronics** ***13***(15): 2998. DOI: 10.3390/electronics13152998.

[6] A. K. Jumani, J. Shi, A. A. Laghari, M. A. Amin, A. U. Nabi, K. Narwani, and Y. Zhang, (2025) *"Quality of Experience (QoE) in Cloud Gaming: A Comparative Analysis of Deep Learning Techniques via Facial Emotions in a Virtual Reality Environment"* **Sensors** ***25***(5): 1594. DOI: 10.3390/s25051594.

[7] A. A. Laghari, S. Shahid, R. Yadav, S. Karim, A. Khan, H. Li, and Y. Shoulin, (2023) *"The State of Art and Review on Video Streaming"* **Journal of High Speed Networks** ***29***(3): 211–236. DOI: 10.3233/JHS-222087.

[8] M. R. Sareddy and V. Kumar. "Virtual Reality Meets AI: Revolutionizing Physical Education with LiDAR and RL". In: *2025 8th International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech)*. IEEE, 2025, 1–6. DOI: 10.1109/IEMENTech65115.2025.10959637.

[9] M. Maciejewski, J. Kocoń, W. Oleksy, M. Kopeć, and W. Kazienko. *amu-cai/CAMEO Dataset*. Accessed: May 9, 2026. 2026.

[10] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. "MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019. DOI: 10.18653/v1/P19-1050.

[11] B. Xie, M. Sidulova, and C. H. Park, (2021) *"Robust Multimodal Emotion Recognition from Conversation with Transformer-Based Crossmodality Fusion"* **Sensors** ***21***(14): 4913. DOI: 10.3390/s21144913.

[12] C. Qu et al., (2025) *"Enhancing Emotion Recognition in Virtual Reality: A Multimodal Dataset and a Temporal Emotion Detector"* **Frontiers in Psychology** ***16***(2): 1709943. DOI: 10.3389/fpsyg.2025.1709943.

[13] H. Wang and M. Wang, (2026) *"Subject-Independent Multimodal Interaction Modeling for Joint Emotion and Immersion Estimation in Virtual Reality"* **Symmetry** ***18***(3): 451. DOI: 10.3390/sym18030451.