MetaPal: A New Paradigm for Human-Computer Interaction — From Command Response to Emotional Empathy

Hui Shao
East China Normal University
71285903006@stu.ecnu.edu.cn

Zizhuo Wu
East China Normal University
71285903022@stu.ecnu.edu.cn

Jing Chen
East China Normal University
51285903016@stu.ecnu.edu.cn

Shiyu Wang
East China Normal University
51285903056@stu.ecnu.edu.cn

Jiaqi Song
East China Normal University
51285903003@stu.ecnu.edu.cn

Yan Wang*
East China Normal University
yanwang@dase.ecnu.edu.cn

ABSTRACT

MetaPal is an intelligent system for cross-cultural empathetic oral dialogue education for international students, applying real-time multimodal affective computing to language education. Unlike traditional "command response" systems, MetaPal takes emotional empathy as the system core across all layers: perception, understanding, decision, and expression. The system fuses visual, audio, and text modalities, driving digital human interaction through an emotion-centered architecture that achieves a paradigm shift from "command response" to "emotional empathy." Through asynchronous processing, model lightweighting, and streaming transmission, the system achieves smooth real-time interaction. Small-scale user trials show positive feedback in oral confidence, cultural awareness, and emotional expression appropriateness. This demo presents the complete technical architecture, interaction process, and preliminary evaluation.

Keywords

Paradigm Shift, Affective Computing, Multimodal Fusion, Digital Human, Cross-Cultural Education

1. INTRODUCTION

1.1 Problem Background

Globalized education has intensified cross-cultural adaptation challenges for international students. Research shows that international students face significant challenges in learning and adjustment during their study abroad experiences [19]. Language learning presents specific challenges:

- **Emotional misunderstanding**: Reserved East Asian expressions are misinterpreted as indifference in Western contexts
- **Non-verbal cues**: Cultural differences in eye contact and facial expression create barriers
- **Vocal emotion**: Native tone transfer makes second language expressions seem emotionally flat

Traditional language tools focus on vocabulary and grammar, ignoring emotional and cultural dimensions. Existing affective computing research remains laboratory-based without systematic integration. Critically, current systems follow the "command response" paradigm, processing speech without perceiving emotional states—emotion is merely optional.

1.2 Related Work

As surveyed in [16], affective computing has established solid foundations for emotion-aware systems. **Application of Affective Computing in Education**: D'Mello et al. [4] analyzed emotional states during learning, showing emotion-cognition synergy is crucial. However, most studies use single modality without real-time multimodal fusion [3, 2]. Picard's foundational work [13] established affective computing as a field, while Pantic and Rothkrantz [11] pioneered affect-sensitive multimodal interaction.

Multimodal Learning Analytics: Existing classroom multimodal analysis frameworks have high latency, limiting real-time interaction support [12, 1]. Blikstein and Worsley [1] demonstrate computational approaches for measuring complex learning tasks.

Cross-Cultural Communication Theory: Hofstede's cultural dimension theory [8] provides a framework for quantifying cultural differences, but few studies have computationalized it for educational systems. Gao et al. [7] examine how cultural epistemologies influence emotional experience and regulation.

* Corresponding author.

Hui Shao, Zizhuo Wu, Jing Chen, Shiyu Wang, Jiaqi Song, and Yan Wang. MetaPal: A New Paradigm for Human-Computer Interaction — From Command Response to Emotional Empathy. In Anthony Botelho, Maria Mercedes T. Rodrigo, Adish Singla, Hiroaki Ogata, Hyojeong So, and Young Hoan Cho (eds.) Proceedings of the 19th International Conference on Educational Data Mining, Seoul, Republic of Korea, June, 2026, pp. 705–709. International Educational Data Mining Society (2026).

https://doi.org/10.5281/zenodo.21040066

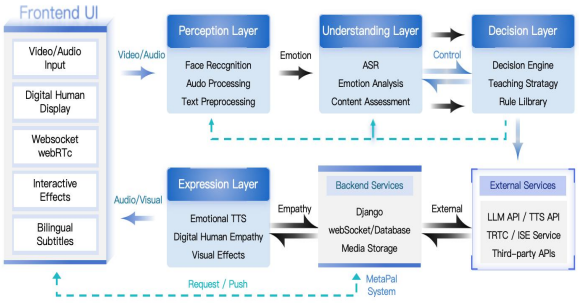

Figure 1: MetaPal architecture: Four-layer framework integrating perception, understanding, decision, and expression layers.

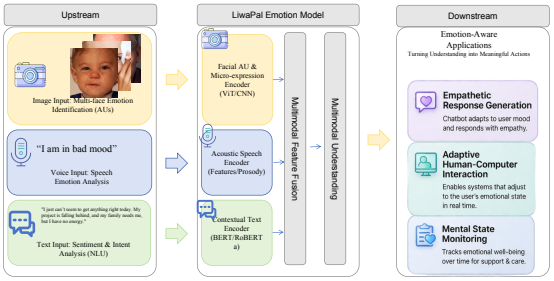

Figure 2: Perception layer: Real-time video capture with picture-in-picture slicing for emotion recognition.

1.3 Research Gap and Contributions

Research Gap: Existing systems lack real-time multimodal fusion, cross-cultural adaptation, and complete engineering implementation. Fundamentally, they follow "command response" logic with emotion as an add-on, while MetaPal takes "emotional empathy" as the operating system core across all layers.

Contributions:

1. **Theoretical**: Propose a new interaction paradigm with emotion as the underlying core—from "command response" to "emotional empathy."
2. **Technical**: Implement emotion-permeating architecture across perception, understanding, decision, and expression layers, with open-source reference.
3. **Practical**: Validate feasibility in real educational scenarios with user feedback and deployment solutions.

Positioning within EDM: While MetaPal draws on HCI and affective computing, its core contribution is **educational data mining**: real-time multimodal affective data as **educational data** to model cross-cultural emotional states in oral dialogue. Unlike prior EDM work on clickstreams or quizzes, MetaPal treats **emotion as a first-class educational data stream** addressing EDM 2026's "Crossing Boundaries" theme.

2. SYSTEM ARCHITECTURE

2.1 Overall Design: Empathetic Architecture

MetaPal implements a four-layer emotion-permeating architecture: perception, understanding, decision, and expression. The perception layer captures multimodal user data for emotion detection. The understanding layer integrates visual, audio, and text features through ViT-based emotion recognition [5], OpenSMILE audio analysis [6], and BERT semantic processing, all centered on emotional comprehension. The decision layer dynamically selects interaction strategies based on emotional states, while the expression layer generates emotionally appropriate responses through digital human feedback (Figure 1).

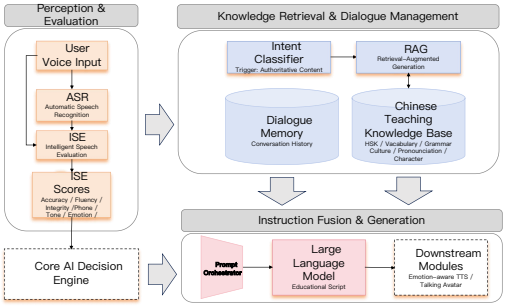

Figure 3: Understanding layer: Three parallel processing paths for multimodal emotional feature extraction and fusion.

2.2 Perception Layer: Emotion Capture

The perception layer captures user video in real-time through camera input, performs picture-in-picture slicing, and transmits the video stream to the backend affective computing module (Figure 2). Following large-scale video-based FER benchmarks [17], our system captures micro-expressions in real-time. Users' expressions, eye contact, and expressions are incorporated into the core processing flow as the starting point of interaction.

2.3 Understanding Layer: Emotion-Semantic Integration

The understanding layer achieves multimodal emotional feature extraction and fusion through three parallel paths: ViT-based visual emotion recognition, OpenSMILE audio emotion analysis, and BERT text semantic analysis (Figure 3). To handle real-world noise, we draw on dual-stage purification methods [14]. Emotion is integrated with semantic parsing—for example, when a user says "I'm fine" with a tense expression, the system interprets it as "tense fine" rather than "calm fine." This approach builds on multimodal affect detection research [3, 10] and affective body expression recognition [9].

2.4 Decision Layer: Emotion-Based Strategy Selection

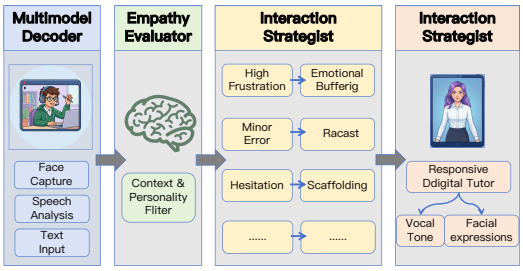

Figure 4: Decision layer: Dynamic strategy selection based on emotional states and learning context.

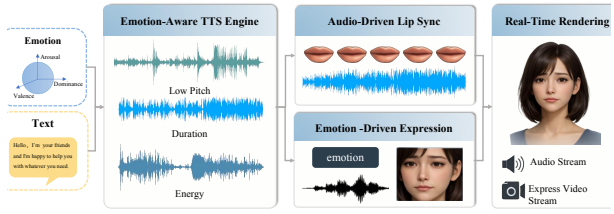

Figure 5: Expression layer: Emotion-driven generation of TTS audio, digital human animations, and visual effects.

The decision engine uses a finite state machine and rule library to dynamically select teaching strategies based on emotional states, cultural background, and learning progress (Figure 4). Emotion forecasting techniques [15] inspire our strategy selection engine. When users are tense, the system slows speech and encourages; when confused, it simplifies and provides scaffolding; when joyful, it increases difficulty and initiates challenges.

Cross-Cultural Parameterization: Hofstede's dimensions [8] are mapped to decision engine weights: power distance, individualism, uncertainty avoidance, and high/low-context communication preferences.

2.5 Expression Layer: Emotion-Driven Output

The expression layer generates TTS audio, digital human animations, and visual effects based on decision results (Figure 5). Audio tone, speed, and rhythm, along with digital human expressions, postures, and interaction methods, are all driven by emotional states to ensure responses are both semantically correct and emotionally appropriate.

3. EVALUATION AND PRELIMINARY FEEDBACK

3.1 Deployment

Figure 6 shows MetaPal deployment in a laboratory environment. Users participate in video calls with the digital human, while the system captures facial expressions in real-time via picture-in-picture for empathetic response.

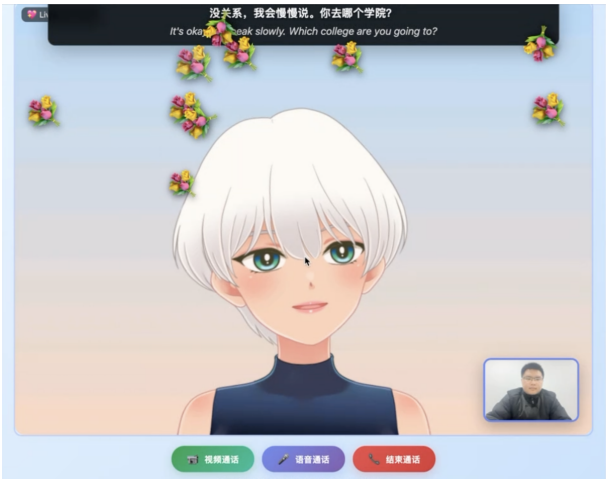

Figure 6: MetaPal deployment: Video calls with real-time emotion capture via picture-in-picture.

3.2 Trial Status

MetaPal is in laboratory prototype verification stage, validated through internal testing and small-scale user trials. The system operated stably with smooth real-time interaction, accurate multimodal emotion recognition, and good audio-visual synchronization.

Participants: International students (N=limited, B1-B2 English) from multiple Asian cultural backgrounds. This diverse sample enables examination of cultural differences in emotional expression [18, 7].

Cross-Cultural Validation: To evaluate the system's cross-cultural adaptation capability, we compared user responses across participants from different cultural backgrounds (Chinese, Korean, Vietnamese). Hofstede's cultural dimension scores [8] for each country were used as reference baselines. Preliminary analysis showed that the system's emotion recognition thresholds appropriately shifted: for example, users from high-context cultures received higher sensitivity weights on facial expression and prosody features, while users from low-context cultures relied more on semantic content analysis. This alignment between system parameters and cultural dimension predictions suggests the feasibility of computationalizing cross-cultural communication theory.

Ethics and Privacy: All trials were conducted with informed consent from participants. As this was a minimal-risk prototyping study, facial expression data were processed in real time and not stored beyond the session; only anonymized behavioral logs were retained for analysis.

Process: Participants completed several weeks of cross-cultural dialogue training with real-time feedback. User experience was collected through questionnaires and interviews, following established methods for evaluating affective educational systems [2].

3.3 Preliminary Results

Small-scale trials showed positive feedback in oral confidence, cultural awareness, and emotional expression. In one case, a user said "Thank you for your help" with a flat expression. The system detected the mismatch and the digital human smiled while explaining gratitude customs in Western culture. User feedback:

> "The system taught me that in Western culture, 'Thank you' needs a smile to be sincere. I wasn't aware of this before."

This demonstrates the paradigm shift—the system "felt" the user's expression and responded emotionally, not just processing the words.

4. DISCUSSION AND CONTRIBUTIONS

4.1 Theoretical Contributions

1. **Paradigm Shift**: MetaPal reconstructs human-computer interaction with emotion as the operating system core across all layers. Command response becomes a subfunction, while emotional empathy drives the interaction, shifting from task-completion to relationship-building.
2. **Emotion-Culture Interaction Model**: Integrates Hofstede's cultural dimension theory [8] into real-time emotion recognition, enabling culturally adaptive expression interpretation.
3. **Multimodal Emotion Fusion**: Verified synergistic effects of visual, audio, and text modalities in language learning.

4.2 Technical Contributions

1. **Emotion-Permeating Architecture**: Implemented four-layer architecture with emotion integrated throughout, providing technical reference for real-time emotional feedback.
2. **Open-source Implementation**: System architecture and core code will be open-sourced on Gitee after acceptance.
3. **Emotional Decision Engine**: Real-time emotion recognition via video calls with dynamic cross-cultural feedback strategies.

4.3 On System Integration as Contribution

We acknowledge that MetaPal builds on established components (ViT, BERT, OpenSMILE, FSM). However, for a **Demo Track** paper, the contribution lies not in proposing new base models, but in demonstrating **how existing technologies can be systematically integrated to solve a real, under-addressed educational problem**. In MetaPal's case, this integration requires: (1) real-time fusion of three modalities under latency constraints, (2) cross-cultural adaptation of emotion recognition thresholds, and (3) driving a digital human's empathetic response. We argue that **system-level innovation**—making working, deployable systems that address genuine educational gaps—is a distinct and valuable form of contribution, complementary to model-level advances, and precisely what EDM's Demo Track is designed to showcase.

4.4 Practical Contributions

1. **Prototype Validation**: Demonstrated technical feasibility in laboratory environment, with real-time multimodal processing under 500ms latency.
2. **Training Solution**: Provided cross-cultural empathetic oral training for international students, addressing emotional expression gaps often overlooked by traditional language tools.
3. **Cross-disciplinary Model**: Demonstrated collaboration across computer science, education, linguistics, and psychology, offering a reusable blueprint for emotion-aware educational systems.
4. **Real-time Fusion Pipeline**: Achieved synchronized processing of video, audio, and text streams with $<$500ms end-to-end latency, enabling natural turn-taking in cross-cultural dialogue.

5. DEMO CONTENT

A demonstration video is submitted as supplementary material, showing live interaction between a user and the MetaPal digital human. The complete system will be open-sourced after acceptance.

6. CONCLUSIONS

MetaPal demonstrates the feasibility of emotion-centered human-computer interaction, providing cross-cultural empathetic oral training for international students. The system bridges computer science and education, with core contributions including:

1. **Paradigm Shift**: Emotion as the operating system core across all layers, redefining interaction from command completion to emotional understanding.
2. **System Completeness**: Complete technical architecture: real-time multimodal perception, emotion-semantic understanding, strategy decision, and empathetic expression generation with full pipeline integration.
3. **Scalability**: Foundation for future large-scale applications with plug-and-play modules for emotion recognition and strategy selection.
4. **Cross-Cultural Adaptability**: Demonstrated real-time parameterization of Hofstede's cultural dimensions for personalized emotion recognition and dialogue strategy adjustment.

Future work will conduct larger-scale experiments to verify educational effectiveness, extend the framework to more cultural contexts (e.g., Latin American, Middle Eastern), and integrate LLM-based open-domain dialogue capabilities.

7. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (No. 62406075), the National Key Research and Development Program of China (2023YFC3604802), and the Shanghai Key Technology R&D Program (Grant No. 25511107200).

8. REFERENCES

[1] P. Blikstein and M. Worsley. Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2):220–238, 2016.

[2] N. Bosch, S. D'Mello, R. S. Baker, J. Ocumpaugh, V. Shute, M. Ventura, X. Wei, and W. Zhao. Automatic detection of learning-centered affective states in the wild. In *International Conference on Intelligent Tutoring Systems*, pages 379–389. Springer, 2016.

[3] S. D'Mello and A. Graesser. Multimodal semi-automated affect detection from conversational cues, gross body language, and facial features. In *User Modeling 2007*, pages 50–59. Springer, 2007.

[4] S. D'Mello and A. Graesser. Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157, 2012.

[5] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR 2021*, 2021.

[6] F. Eyben et al. Opensmile: The munich versatile and fast open-source audio feature extractor. In *ACM Multimedia*, pages 1459–1462, 2010.

[7] G. Gao, J. Spencer-Rodgers, and L. Ku. Dialectical emotions: How cultural epistemologies influence the experience and regulation of emotion. *Social and Personality Psychology Compass*, 7(11):838–854, 2013.

[8] G. Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1):8, 2011.

[9] A. Kleinsmith and N. Bianchi-Berthouze. Affective body expression perception and recognition: A survey. *IEEE Transactions on Affective Computing*, 4(1):15–33, 2013.

[10] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 2004(11):1672–1687, 2004.

[11] M. Pantic and L. J. M. Rothkrantz. Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE*, 91(9):1370–1390, 2003.

[12] R. Pekrun. The control-value theory of achievement emotions. *Educational Psychology Review*, 18(4):315–341, 2006.

[13] R. W. Picard. Affective computing: Challenges. *International Journal of Social Robotics*, 2(3):193–198, 2010.

[14] H. Wang, X. Mai, Z. Tao, X. Tong, J. Lin, Y. Wang, J. Yu, S. Yan, Z. Zhou, and W. Zhang. D2sp: Dynamic dual-stage purification framework for dual noise mitigation in vision-based affective recognition. In *CVPR*, 2025.

[15] H. Wang, X. Mai, Z. Tao, J. Yu, Z. Zhou, X. Tong, S. Yan, Q. Zhao, S. Gao, Y. Wang, and W. Zhang. Hi-ef: Benchmarking for human-interaction-based emotion forecasting. In *AAAI*, 2026.

[16] Y. Wang, W. Song, W. Tao, D. Yang, A. Liotta, X. Li, S. Gao, Y. Sun, W. Ge, W. Zhang, and W. Zhang. A systematic review on affective computing: Emotion models, databases, and recent advances. *Information Fusion*, 2022. ESI Highly Cited Paper.

[17] Y. Wang, Y. Sun, Y. Huang, Z. Liu, S. Gao, W. Zhang, W. Ge, and W. Zhang. Ferv39k: A large-scale multi-scene dataset for facial expression recognition in videos. In *CVPR*, 2022.

[18] A. E. Woolfolk, K. A. Davis, and S. J. Pape. Teacher talk to diverse learners: A study of teacher-student communication in regular classrooms. *Teaching and Teacher Education*, 22(5):536–546, 2006.

[19] Y. Zhang and A. Goodwin. Unpacking the complexity of international chinese doctoral students' learning and adjustment experiences. *Higher Education*, 62(5):595–613, 2011.