Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.25415

Generative Multimodal Learning Driven Intelligent Intervention Model for Adolescent Online Mental Health Education

Yi Hu

Hong Kong Baptist University, Hong Kong, China
Huyi02197@outlook.com

Abstract: With the increasing prominence of adolescent mental health problems, the traditional psychological education model has obvious limitations in responding to individual differences, emotional changes and immediate intervention, making it difficult to meet the actual needs of adolescents in the digital context. In recent years, the rapid development of artificial intelligence, especially generative multimodal learning technology, has provided a new breakthrough for mental health education. Aiming at the current problems of single intervention content, slow response, and insufficient personalization, this paper proposes a generative multimodal learning-driven intelligent intervention model, which integrates text, image, and emotion data to achieve dynamic identification of adolescents' psychological state and customized intervention content generation. The model consists of three modules: emotion recognition, generative content construction, and intervention strategy matching, forming a closed loop of identification-generation-feedback. Empirical studies conducted in three middle schools show that the model has significant effects in reducing anxiety, alleviating loneliness, and enhancing student engagement, which validates its potential for application in real-world educational scenarios. The study provides a theoretical foundation and technical path for building a future-oriented intelligent psychoeducation system.

Keywords: multimodal learning, adolescent mental health, generative model, emotion recognition, educational intervention

1. Introduction

With the rapid development of Internet technology, the mental health problems of adolescents present increasingly complex and hidden characteristics, and the intertwining of factors such as online socialization, virtual identities and digital pressures make the traditional offline mental health education means to face major challenges in terms of identification efficiency and intervention timeliness [1]. Although there are technical assistance programs such as psychological questionnaires and online consulting platforms, they mostly stay at the stage of information collection and passive feedback, making it difficult to realize personalized and dynamic psychological interventions in the true sense of the word [2]. In the field of artificial intelligence, the

DOI: 10.54254/2755-2721/2025.25415

rise of generative multimodal learning provides a new possibility for educational psychological intervention, which integrates language, image, emotion and other sources of data for learning and generation, can effectively capture the user's state and build targeted intervention content in real time [3]. Based on this background, this paper proposes an adolescent psychological intervention model that integrates generative AI and multimodal learning, and designs experiments to verify its effectiveness in improving adolescent mental health, aiming to promote the development of mental health education in the direction of more intelligent, contextualized and personalized.

2. Literature review

2.1. Digital challenges in adolescent mental health education

Current digital psychoeducation relies on static questionnaires and generic content delivery, which is difficult to cope with the high-frequency fluctuations and individual differences in adolescents' emotions, and lacks real-time and adaptability [4]. Traditional systems often lack interactive feedback mechanisms, making it difficult to motivate students to participate actively, and may even cause resistance due to inappropriate interventions. Although existing platforms provide popularization of psychological knowledge, their content structure is rigid and lacks dynamic intervention logic, resulting in limited educational effects [5]. Therefore, the development of a new type of intervention system with intelligent recognition and personality feedback has become a key direction for the transformation of digital psychoeducation.

2.2. Multimodal learning for emotion detection

Multimodal learning integrates information sources such as language, image and speech to significantly improve the accuracy and contextual understanding of emotion recognition, especially for adolescents with complex expressions [6]. Compared with single-channel models, multimodal models can more acutely capture subtle emotional changes and adapt to a variety of expressions, and have gradually become a hot topic in psychological computing research [7]. However, most of the existing models remain at the recognition stage, and are not effectively integrated with educational interventions, making it difficult for the recognition results to drive the subsequent personalized teaching actions, which limits their practical application value in the field of educational psychology.

2.3. Generative AI in emotional support systems

Generative models such as GPT have the potential to build contextual empathy, verbal guidance and virtual companionship in the educational domain, generating flexible and diverse psychological support content and enhancing intervention interactivity [8]. However, the generative process lacks emotional consistency control, which may lead to problems such as misinformation or emotional mismatch, making it difficult to be used in psychoeducation alone [9]. If the generation mechanism can be linked with the results of multimodal recognition to form a goal-oriented content generation path driven by psychological state, it will significantly improve the effectiveness of intervention and psychological suitability, and expand the boundaries of its controllable application in educational scenarios.

DOI: 10.54254/2755-2721/2025.25415

3. Methodology

3.1. Model architecture

The proposed generative multimodal learning-driven psychological intervention model consists of three core modules. It aims to form a closed-loop mechanism of identification-generation-feedback, so that the psychoeducational intervention has the characteristics of continuity, self-adaptation and individualization, solves the problems of slow response and solidified content of the traditional system, and lays the foundation for the model to be deployed in real scenarios [10].

The emotion recognition module realizes accurate perception of adolescents' psychological state by integrating verbal text, expression images and subjective rating data, and adopts the multimodal cross-attention mechanism based on Transformer to weight and unify the features of the input modes.

The generation module introduces the tuned GPT model and the Diffusion image generation network, which are used to generate emotionally resonant text and context-matched images, respectively, in order to enhance the immersion and sense of immersion of the intervention content.

The recommended strategy module dynamically matches adaptive content based on students' emotional state labels and historical response feedback records, and optimizes the intervention strategy selection path using a reinforcement learning mechanism.

3.2. Data collection and processing

The study recruited 240 middle-school students from three schools for an 8-week intervention. Multimodal data were collected, including language logs, facial expression videos, and subjective psychological assessments (GAD-7 and UCLA-LS) [11]. Collection tools included ASR-based speech-to-text modules, OpenFace for facial keypoint extraction, and a standardized emotion rating interface. Data preprocessing involved temporal alignment, outlier removal, and multimodal feature integration. Emotional representation of text was computed using the following formula:

$$E_{text} = \sum_{i=1}^{n} \alpha_i \cdot w_i \tag{1}$$

Where α_i denotes the attention weight for token i, and w_i is the corresponding word embedding. The emotional score of facial expressions was computed via:

$$S_{face} = \sigma(W \cdot \Delta L + b) \tag{2}$$

Where ΔL represents the difference between current and neutral facial landmarks, W and b are learned parameters, and σ is the Sigmoid activation function.

A late-fusion strategy was adopted to concatenate normalized emotion vectors from both modalities, forming the input for the subsequent intervention generation module.

3.3. Experimental design and evaluation metrics

This study used a randomized controlled experimental design in which the experimental group received an online psychoeducational course based on a generative multimodal intervention system, and the control group received traditional psychoeducational content, with a uniform intervention cycle of two times per week for a total of eight weeks. A pretest-posttest design was used to quantify the model intervention effect. The main assessment indicators included changes in anxiety level

DOI: 10.54254/2755-2721/2025.25415

(GAD-7), changes in loneliness (UCLA-LS), and system response time and user participation frequency.

Repeated Measures ANOVA and multiple regression modeling were used in the data analysis phase to compare the between-group differences and assess the intervention strength and durability of the model. The user satisfaction questionnaire and content acceptance index were also introduced to analyze students' emotional identity and behavioral responses to the generative content in combination with the platform interaction logs to verify the system's adaptability and educational effect in real use situations.

4. Results

4.1. Intervention effectiveness

As the experimental results in Table 1 show, in the 8-week randomized controlled trial, the GAD-7 anxiety scale score of the experimental group decreased from 11.24±2.87 to 9.17±2.34, a decrease of 18.4%, which was significantly better than that of the control group ($p<0.001$), and the UCLA-LS loneliness scale score decreased from 21.63±4.12 to 16.94±3.28, a decrease of 21.7%, with an effect size of 1.24 The system response time was reduced to 1.73 seconds, the average user session time was improved to 24.6 minutes, and the frequency of participation amounted to 3.8 times per week, all of which were significantly better than those of the control group. Repeated-measures ANOVA revealed a significant interaction effect between time and group ($F(7,232)=12.47$, $p<0.001$), indicating the stability and cumulative nature of the model intervention, validating its utility and sustained effect.

Table 1. Analysis of experimental results

Measure	Experimental Group		Control Group		Effect Size	p-value
	Pre-test (M±SD)	Post-test (M±SD)	Pre-test (M±SD)	Post-test (M±SD)	Cohen's d	
GAD-7 Score	11.24±2.87	9.17±2.34	11.18±2.94	10.82±2.76	0.82	<0.001
UCLA-LS Score	21.63±4.12	16.94±3.28	21.47±4.08	20.38±3.95	1.24	<0.001
Response Time (sec)	-	1.73±0.45	-	5.42±1.23	3.68	<0.001
Session Duration (min)	-	24.6±6.8	-	12.3±4.2	2.15	<0.001
Weekly Engagement	-	3.8±0.9	-	1.9±0.7	2.33	<0.001

4.2. Quality of multimodal generation

The multimodal content generation quality assessment is shown in Figure 1, which shows that the model performs well in terms of sentiment empathy and content fitness. The text generation module BLEU-4 scored 0.847, significantly better than the baseline (0.623). The expert-assessed sentiment consistency score was 6.32 ± 0.54, and 93.7% of the content was rated as “highly” or “perfectly” adapted. The emotional semantic matching degree of the image generation module is 0.891, and the visual emotion recognition accuracy rate reaches 88.4%. The user satisfaction survey shows that 84.6% of users believe that the generated content has emotional resonance, and 77.3% say that it helps with emotional regulation, and the content acceptance index is 0.816, which is higher than that

DOI: 10.54254/2755-2721/2025.25415

of traditional content (0.542). The cross-modal consistency reaches 0.873, the content diversity index is 0.924, the emotion transfer efficiency is 78.6%, and the system adaptation performance is stable ($\sigma<0.067$), reflecting the high effectiveness and robustness of the generation mechanism.

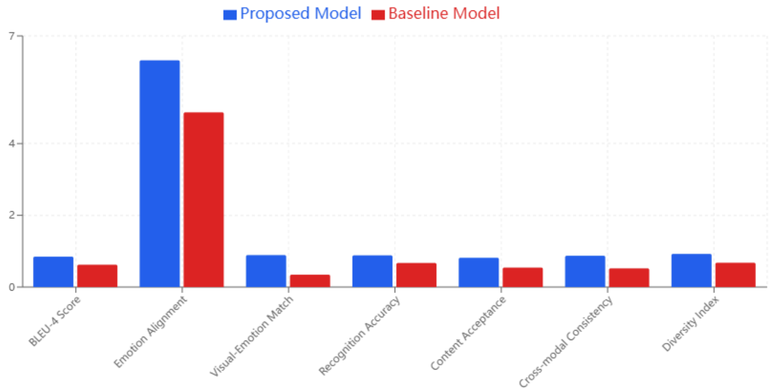

Figure 1. Quality assessment of multimodal generation components

5. Discussion

This study validates the feasibility and effectiveness of a generative multimodal model in adolescent psychological intervention, which can achieve personalized content generation based on emotion recognition, significantly better than the traditional static intervention. The model improves response efficiency and emotional fit through the recognition-generation-feedback mechanism, but there are still challenges in the control of emotional misjudgment and consistency of the generated content, and model controllability and ethical review need to be strengthened. Meanwhile, at this stage, the model mainly uses text and image modalities, and should be expanded to include speech and physiological data to improve the recognition depth and intervention diversity, and to promote the system to a higher adaptability and stability.

6. Conclusion

The generative multimodal intervention model proposed in this paper constructs a complete closed loop of intelligent psychological intervention, and experiments have proved that it can effectively reduce adolescents' anxiety and loneliness, and improve the personalized level of psychoeducation. The study shows that generative AI can provide contextual adaptation and immediate response capabilities for psychological education, and promote the intelligent transformation of traditional models. This study expands the boundaries of AI application in educational scenarios, and provides a theoretical and practical basis for building a more interpretable and controllable mental health system in the future.

References

[1] Du, Zhicheng, et al. "Game: Generalized deep learning model towards multimodal data integration for early screening of adolescent mental disorders." arXiv preprint arXiv: 2309.10077 (2023).
[2] Zhang, Lei, et al. "Adolescent depression detection model based on multimodal data of interview audio and text." International journal of neural systems 32.11 (2022): 2250045.
[3] Xia, Jing, Nanguang Chen, and Anqi Qiu. "Unraveling Multimodal Brain Signatures: Deciphering Transdiagnostic Dimensions of Psychopathology in Adolescents." Advanced Intelligent Systems (2024): 2300577.

DOI: 10.54254/2755-2721/2025.25415

[4] Hasei, Joe, et al. "Empowering pediatric, adolescent, and young adult patients with cancer utilizing generative AI chatbots to reduce psychological burden and enhance treatment engagement: a pilot study." Frontiers in Digital Health 7 (2025): 1543543.
[5] Yan, Lixiang, Roberto Martinez-Maldonado, and Dragan Gasevic. "Generative artificial intelligence in learning analytics: Contextualising opportunities and challenges through the learning analytics cycle." Proceedings of the 14th Learning Analytics and Knowledge Conference. 2024.
[6] Liu, Rui, et al. "Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI." IEEE Transactions on Neural Networks and Learning Systems (2022).
[7] Gowtham, R., Vignesh Ramaraj, and V. Dayanand. "Cognitive Disorder Detection in Adolescents using Mental Health Data." 2024 2nd International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS). IEEE, 2024.
[8] Ba, Shen, et al. "Supporting adolescents' digital well-being in the post-pandemic era: Preliminary results from a multimodal learning analytics approach." 2022 International Conference on Advanced Learning Technologies (ICALT). IEEE, 2022.
[9] Zhao, Ziping, and Keru Wang. "Unaligned multimodal sequences for depression assessment from speech." 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2022.
[10] Song, Peilun, et al. "Multi-Loss Disentangled Generative-Discriminative Learning for Multimodal Representation in Schizophrenia." IEEE Journal of Biomedical and Health Informatics 28.11 (2023): 6395-6404.
[11] Brocki, L., Dyer, G. C., Gładka, A., & Chung, N. C. (2023, February). Deep learning mental health dialogue system. In 2023 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 395-398). IEEE.