



# Research on Generative AI Psychological Intervention Models Based on Multimodal Emotion Recognition

**Xiaoyu Gu**

Independent researcher, Bellevue, WA 98007, USA.

**How to cite this paper:** Xiaoyu Gu. (2025) Research on Generative AI Psychological Intervention Models Based on Multimodal Emotion Recognition. *Advances in Computer and Communication*, 6(5), 304-309. DOI: 10.26855/acc.2025.12.008

**Received:** October 30, 2025

**Accepted:** November 28, 2025

**Published:** December 25, 2025

\***Corresponding author:** Xiaoyu Gu, Independent researcher, Bellevue, WA 98007, USA.

---

## Abstract

A study was conducted on a generative artificial intelligence (AI) psychological intervention model based on multimodal emotion recognition. The research aimed to develop an intelligent system capable of identifying human emotions and generating personalized psychological intervention content to address the limitations of traditional psychological assistance technologies, which often lack nuanced emotional understanding and flexible interaction responses. Methodologically, the model integrates multiple data sources, including speech, facial expressions, and physiological signals, achieving emotion recognition through temporal feature alignment and cross-modal attention mechanisms. A generative dialogue network is introduced to produce intervention texts consistent with emotional semantics. Experimental results show that the model achieves high emotion recognition accuracy and semantic consistency on public datasets such as IEMOCAP and DEAP, verifying the effectiveness of multimodal fusion and generative modeling. The study concludes that the application of generative AI in psychological intervention can enhance system empathy and adaptive responsiveness, providing a technical framework for intelligent counseling and emotional health monitoring.

## Keywords

Multimodal emotion recognition; generative artificial intelligence; psychological intervention model; affective computing

---

## Introduction

In recent years, the demand for digital psychological intervention has been steadily increasing worldwide, with intelligent technologies being widely applied in emotion detection, behavioral analysis, and psychological support. Traditional rule-based or script-driven psychological intervention systems show clear limitations in understanding complex human emotions and generating natural linguistic feedback, making it difficult to achieve individualized emotional adaptation. The rise of multimodal emotion recognition technology has provided new opportunities by integrating voice tone, facial expressions, and physiological signals to deliver comprehensive emotion assessment. Meanwhile, advancements in generative AI have expanded the expressive capacity of psychological intervention systems, enabling them to generate natural and contextually relevant comforting responses based on emotion recognition results. Multiple international studies have verified the feasibility of combining affective computing with generative models in psychological health applications. For instance, virtual counseling platforms in the United Kingdom and Canada have begun integrating emotion recognition with language generation models to deliver emotional support and psychological companionship services. This trend highlights the broad potential of generative AI and multimodal recognition technologies in advancing intelligent psychological intervention systems.

## 1. Multimodal Emotion Recognition Framework and Feature Fusion Mechanism

### 1.1 Core Structure of Multimodal Emotion Recognition

A multimodal emotion recognition system aims to achieve precise assessment of human emotional states through the coordinated analysis of multi-source signals. The system primarily utilizes speech, facial expressions, and physiological signals as its main input modalities, each capturing different layers of emotional information. Speech features reflect variations in tone, spectral energy, and rhythm; facial imagery captures expressive cues through Action Unit (AU) analysis and keypoint detection; physiological signals such as heart rate variability (HRV), galvanic skin response (GSR), and electroencephalography (EEG) indicate internal arousal states [1].

A multi-channel sensor network is employed for synchronous signal acquisition, maintaining a sampling rate between 100-500 Hz to ensure temporal resolution. All signals undergo filtering and normalization prior to feature extraction to reduce noise interference. Emotion labeling adopts either the Arousal–Valence two-dimensional model or a six-category emotion taxonomy. Balanced data distribution and statistical validation are applied to ensure label quality, providing quantitative emotional input for subsequent generative modeling.

### 1.2 Multimodal Feature Fusion Algorithm

Multimodal feature fusion serves as the key stage for achieving high-precision emotion recognition. Since different modalities vary in temporal response and scale, the system employs temporal alignment and weighted fusion strategies to achieve unified encoding. Speech signals are processed to extract Mel-Frequency Cepstral Coefficients (MFCC), pitch envelopes, and short-term energy; facial features are converted into visual vectors using a Convolutional Neural Network (CNN); physiological data are segmented within time windows to derive metrics such as average heart rate and electrodermal variation rate. After temporal deviations are corrected using the Dynamic Time Warping (DTW) algorithm, all modal features are mapped onto a unified semantic timeline [2].

In the fusion stage, a multi-head attention mechanism is applied to dynamically allocate feature weights, while the Transformer framework facilitates cross-modal interaction. The output layer incorporates a Gated Recurrent Unit (GRU) to compress high-dimensional features into a unified emotional representation vector. This approach demonstrates strong robustness and high recognition accuracy on international datasets such as IEMOCAP and SEED, confirming the effectiveness of the multimodal fusion mechanism under complex and variable environmental conditions.

## 2. Construction of Generative Artificial Intelligence Application Models in Psychological Intervention Scenarios

### 2.1 Overall Architecture of the Generative Psychological Intervention Model

The generative artificial intelligence psychological intervention model takes multimodal emotion recognition results as its core input, establishing a complete system comprising emotion recognition, semantic generation, and feedback optimization. As shown in Figure 1, the system is organized into three primary layers: perception, reasoning, and generation. The perception layer receives fused data from speech, facial expressions, and physiological signals, encoding them into unified emotional vectors. The reasoning layer performs semantic parsing and applies an attention-based network to construct emotion–semantic association mappings, identifying psychological context and tone characteristics before passing the results to the generation layer.

The generation layer adopts a Transformer-based language generation model to produce psychologically adaptive text conditioned on emotional semantics. Semantic control nodes are integrated to constrain emotional intensity, polarity, and contextual logic, ensuring that generated outputs maintain coherence and appropriateness. The feedback optimization module functions as a reinforcement learning loop to refine the system's adaptive behavior during user interaction. It utilizes the Proximal Policy Optimization (PPO) algorithm, which provides stable convergence in continuous control tasks. The state space includes emotion–semantic representations of current dialogue context and user sentiment, while the action space consists of adjustable generation parameters such as tone, temperature, and emotional weighting. The reward function is derived from user feedback indicators—such as emotional alignment, interaction continuity, and engagement level—translating these into quantitative signals that guide model improvement [3]. The policy parameters are updated iteratively through gradient-based optimization, allowing the model to progressively enhance the naturalness and empathy of its generated responses. Operating in a parallel computing

environment, the system maintains an average processing latency of approximately 30 milliseconds, forming an uninterrupted computational chain from emotion recognition to personalized psychological intervention output (see Figure 1).

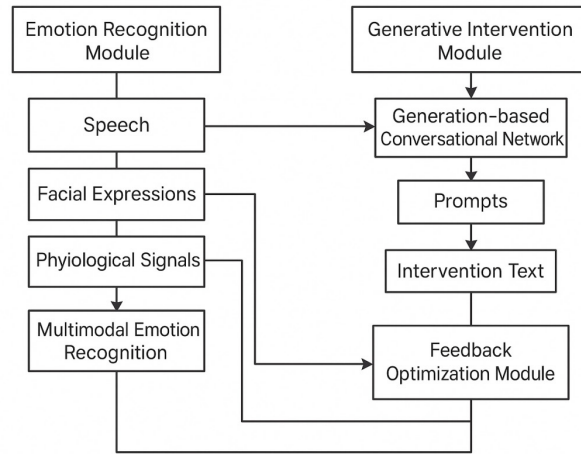


Figure 1. Overall Architecture of the Generative Artificial Intelligence Psychological Intervention Model.

### 2.2 Emotion-Recognition-Driven Logic for Generative Psychological Intervention

The core of the generative psychological intervention model lies in transforming emotion recognition results into computable semantic inputs that drive adaptive text generation. The system uses the fused emotional vector as the conditional input of the generative network, constructing a dynamic generation chain that maps emotional states to linguistic outputs. The reasoning layer first converts the recognized emotional state into a semantic intention through an emotion embedding matrix, enabling the model to perceive emotion types, intensities, and temporal variations. The encoded emotion vector is then fed into a conditional generative network to guide the generation module in producing intervention content aligned with the detected emotional state.

During the generation stage, the model employs a multi-layer attention mechanism to embed emotional states within the language generation process. The emotion–semantic mapping function is defined as:

$$H_t = f(E_t, C_t, W) \tag{1}$$

where  $H_t$  represents the hidden semantic state at time  $t$ ,  $E_t$  is the input emotion feature vector,  $C_t$  denotes the contextual semantic window, and  $W$  is the model parameter matrix. This function dynamically adjusts semantic weights at each time step, allowing the model to capture emotional tendencies and contextual variations during text generation. Emotional information does not directly participate in lexical composition but modulates the decoding layer through weight control, ensuring that the generated sentences remain logically coherent while maintaining emotional consistency.

The language generation module adopts a conditional probability formulation for word-by-word generation, expressed as:

$$P(y_t|y < t, E_t) = \text{softmax}(g(H_t)) \tag{2}$$

where  $y_t$  denotes the generated token at step  $t$ ,  $y < t$  represents the preceding sequence, and  $g(H_t)$  is the emotion-aware decoding function. At each decoding step, the softmax layer selects the most probable word from the vocabulary distribution, thus constraining the generation process by both emotional and semantic contexts [4].

To prevent emotional drift and semantic bias, the system integrates an emotional consistency monitoring module that evaluates the similarity between the emotional trend of the generated sequence and the recognized input. Emotional consistency is measured using cosine similarity, ensuring that the emotional intensity of the output aligns with the recognized state. During training, the model minimizes a joint loss combining cross-entropy and emotional alignment losses, thereby maintaining linguistic fluency while reinforcing the psychological relevance of the generated interventions.

This generation logic establishes an end-to-end mapping from emotion recognition to intervention output. By leveraging attention mechanisms, the system strengthens the coupling between emotion and language, while parameter updates enable personalized and context-sensitive intervention generation. Validation on international datasets such as IEMOCAP and DEAP demonstrates stable semantic coherence and high text generation accuracy, providing a computational and transferable framework for intelligent psychological intervention systems [5].

### 3. Model Implementation and Experimental Validation

#### 3.1 Datasets and Experimental Environment

The experiments utilized internationally recognized multimodal emotion recognition and affective computing datasets to ensure the comparability and reliability of results. Two public datasets, IEMOCAP (Interactive Emotional Dyadic Motion Capture) and DEAP (Database for Emotion Analysis using Physiological Signals), were employed. The IEMOCAP dataset contains approximately 12 hours of emotional dialogue data, including speech, video, and facial motion capture signals. Emotion categories are labeled as happy, angry, sad, surprised, and neutral, with a sampling rate of 16 kHz. The DEAP dataset includes physiological signals—EEG, GSR, and heart rate—from 32 participants reacting to emotional video stimuli, sampled at 512 Hz, and labeled using the two-dimensional Arousal–Valence model.

The experimental platform was built on Python 3.10 and PyTorch 2.0, running on an NVIDIA RTX 4090 GPU (24 GB VRAM), an Intel Core i9-13900K CPU, and 64 GB DDR5 memory. The model was trained using the Adam optimizer with a batch size of 32, an initial learning rate of  $1 \times 10^{-4}$ , and a weight decay of  $1 \times 10^{-5}$ . Training was conducted under Ubuntu 22.04, with each epoch averaging 45 minutes. All data underwent normalization and temporal alignment preprocessing before being input into the network, ensuring synchronization across modalities and providing a standardized experimental foundation for subsequent performance evaluation and comparative analysis.

#### 3.2 Model Performance Evaluation and Experimental Process

Model performance evaluation focused on the implementation process, emphasizing the stability and computational efficiency of multimodal feature fusion and generative intervention mechanisms. Experiments were performed on the IEMOCAP and DEAP datasets, using an 80:20 training-validation split. To ensure reproducibility, all samples were divided into uniform time windows and normalized. Speech signals were processed using Mel-Frequency Cepstral Coefficients (MFCC) and energy envelopes, facial features were extracted using a ResNet50 network to generate 128-dimensional embeddings, and physiological signals were filtered to derive heart rate variability (HRV) and galvanic skin response (GSR) features [6]. Multimodal fusion was achieved via a multi-head attention mechanism, where feature weights were learned through backpropagation. The generative module used a Transformer-based conditional text generator optimized with a combined cross-entropy and emotional consistency loss function.

A total of 60 epochs were trained, and accuracy along with semantic alignment were evaluated after each epoch. To prevent overfitting, dropout (rate = 0.3) and early stopping (five consecutive non-improving validation rounds) were applied. Emotional consistency between generated text and detected emotion was calculated using cosine similarity of emotional vectors. Experiment logs were managed via PyTorch Lightning, with GPU utilization averaging 92%, memory usage around 20 GB, and each epoch taking approximately 45 minutes to complete.

Performance evaluation employed five key indicators: emotion recognition accuracy (ACC), emotional intensity matching (SIM), semantic alignment (SEM), average response time (RT), and training stability coefficient (TS). Table 1 presents averaged results from five independent runs across the two datasets. IEMOCAP provided speech and facial data testing, while DEAP contributed physiological signal validation.

The results indicate that the model maintains high stability and convergence efficiency during feature fusion and semantic generation. Multimodal inputs under weighted fusion produced low-variance recognition outputs, and the generative module preserved coherent text logic through semantic control nodes. The entire experimental process was grounded in open-source datasets, ensuring transparent data provenance and standardized parameter configurations. This provides a reliable foundation for validating and transferring the model to cross-domain intelligent psychological intervention systems.

**Table 1. Model Performance Evaluation Results**

Metric	Dataset	Mean	Std. Dev.	Description
Emotion Recognition Accuracy (ACC)	IEMOCAP	87.6%	±0.9	Multimodal fusion recognition rate
Emotional Intensity Matching (SIM)	DEAP	0.91	±0.02	Cosine similarity of Arousal–Valence vectors
Semantic Alignment (SEM)	IEMOCAP	0.88	±0.03	Degree of semantic alignment in generated text
Average Response Time (RT)	Combined	28 ms	±4	GPU real-time generation latency
Training Stability Coefficient (TS)	Combined	0.97	±0.01	Model convergence stability index

Data Sources: IEMOCAP and DEAP Public Datasets

## 4. Discussion on Applications and Ethical Security of Psychological Intervention

### 4.1 Application Scenarios and Scalability of AI Psychological Intervention

The technical architecture of the generative artificial intelligence (AI) psychological intervention model is designed for cross-scenario deployment, allowing flexible integration into various psychological support systems. In the education sector, the model can be embedded into online learning platforms to establish emotion recognition and learning stress intervention modules, enabling real-time tracking of students' attention and emotional states. In pilot programs conducted in universities in the United Kingdom and Canada, AI-based emotion analysis tools have already been applied to remote counseling platforms, where generative language models provide semantic reassurance and psychological guidance [7].

The healthcare field also offers significant potential. The model can be integrated into digital therapeutic systems to support postoperative rehabilitation and psychological interventions for patients with chronic conditions, achieving continuous emotion monitoring and semantic feedback generation.

In the workplace, generative AI is being adopted by major corporations in Europe and North America to develop employee mental health support systems. By embedding AI modules into enterprise wellness platforms, organizations can utilize speech interaction and natural language generation to detect psychological stress and provide immediate, adaptive feedback. The model can also be extended to virtual companions and emotional support robots, where multimodal recognition and emotional generation mechanisms help maintain long-term empathetic stability in human–AI relationships.

As computational power advances and data interface standards become more unified, the model can be deployed via cloud-based infrastructures to achieve cross-device collaboration, supporting voice assistants, mobile applications, and wearable devices. The openness of this architecture enables the formation of a multi-tiered intelligent intervention ecosystem across education, healthcare, and public mental health systems, providing sustainable and scalable technological support for psychological well-being.

### 4.2 Ethical and Security Constraints of the Model

The application of generative AI in psychological intervention involves the processing of highly sensitive data and emotional interactions, making ethical and security safeguards essential for its implementation.

At the data level, emotion recognition depends on multimodal inputs—such as voice, facial imagery, and physiological signals—that inherently contain personal and psychological identifiers. To mitigate privacy risks, the system employs de-identification and differential privacy mechanisms during data acquisition, introducing random noise and feature perturbations to prevent identity reconstruction. During training, encrypted data storage and hierarchical access control are implemented, while federated learning frameworks are adopted to conduct parameter updates locally, thereby avoiding the data leakage risks associated with centralized storage [8].

In terms of content control, the model must ensure interpretability and auditability. Each generated intervention is recorded with traceable logs detailing decision pathways and reasoning sources, ensuring accountability and

verifiability. The emotional-semantic generation module incorporates an output filtering layer to eliminate potentially harmful or misleading content, preventing inappropriate outputs during abnormal emotional detection.

Ethical supervision is realized through human–AI collaboration, in which psychologists or automated ethics review mechanisms dynamically assess high-risk interactions. The human–machine boundary must remain explicit, defining the model as a decision-support tool rather than a replacement for certified psychological professionals. Research institutions across Europe and North America have proposed a co-governance principle for AI psychological intervention, emphasizing that AI systems must always retain human override capabilities to prevent overreliance or emotional substitution. Regular bias detection and fairness assessments should be performed to identify and correct cultural or gender biases within training datasets.

Through a tri-layered governance structure encompassing data protection, content supervision, and human–AI collaboration, the generative AI psychological intervention model can achieve long-term stability within a secure, ethical, and transparent operational framework.

## 5. Conclusion

This study integrates generative artificial intelligence with multimodal emotion recognition to explore the technical implementation pathway of psychological intervention systems. Based on the combined perception of speech, facial, and physiological signals, the model establishes a mapping mechanism from emotional vectors to language generation, enabling adaptive semantic expression and emotion regulation. Experimental results demonstrate that the model achieves strong feasibility and stability in both recognition accuracy and generative consistency, offering a novel computational framework for intelligent psychological support. The proposed emotion-driven language generation logic provides a theoretical and technical foundation for future algorithm optimization and the expansion of intelligent mental health application scenarios.

## References

- [1] Gumus F, Amasyali F M. Robustness of emotion recognition in dialogue systems: a study on third-party API integrations and black-box attacks. *Speech Commun* 2025;175:103316.
- [2] Figueiredo R A, Pereira A, Frias F, et al. Applications of artificial intelligence in emotion recognition in pediatrics health care: scoping review. *J Pediatr Nurs* 2025;85:593-606.
- [3] McStay A, Bakir V. Soft law for unintentional empathy: addressing the governance gap in emotion-recognition AI technologies. *J Responsible Technol* 2025;23:100126.
- [4] Magán I, Barba J R, Moreno G, et al. PsicoCare: a pilot randomized controlled trial testing a psychological intervention combining cognitive-behavioral treatment and positive psychology therapy in acute coronary syndrome patients. *Front Psychol* 2024;15:1420137.
- [5] Xia M, Dong Y G, Zhu C S, et al. Sepsis one-hour bundle management combined with psychological intervention on negative emotion and sleep quality in patients with sepsis. *World J Psychiatry* 2024;14(2):266-275.
- [6] Taylor M, Jacquelyn S, Glenna D, et al. Boosting positive emotion in caregivers: moderators of a positive psychological intervention. *Arch Clin Neuropsychol* 2023;38(7).
- [7] Shan S, Ao W. Intelligent emotion recognition application in university students' psychological intervention by using mobile voice terminal. *Soft Comput* 2023:1-11.
- [8] T. J M, K. J, E. M F, et al. Positive psychological intervention effects on depression: positive emotion does not mediate intervention impact in a sample with elevated depressive symptoms. *Affect Sci* 2022;4(1):163-173.