Towards Advanced Emotional Care: Embodied Emotional Care System for Humanoid Robots

Yang Chang[1], **Aoxing Li**[1], **Yuxuan Lin**[2], **Jianan Wang**[1], **Lizheng Liu**[1], **Yang Liu**[4], **Jing Liu**[3], **Liang Cao**[5], **Yan Wang**[1,*], **Zhongxue Gan**[1,*], **Wenqiang Zhang**[1,2,*]

[1]Academy for Engineering & Technology, Fudan University, Shanghai, China
[2]School of Computer Science, Fudan University, Shanghai, China
[3]School of Information Science and Technology, Fudan University, Shanghai, China
[4]Department of Computer Science, The University of Toronto, Toronto, Canada
[5]Department of Chemical Engineering, Massachusetts Institute of Technology, Boston, United States
Emails: {ychang24, axli24, yuxuanlin24, wangjn24}@m.fudan.edu.cn, yangliu@cs.toronto.edu, liangcao@mit.edu, {lzliu, jingliu19, yanwang19, ganzhongxue, wqzhang}@fudan.edu.cn

2025 IEEE International Conference on Multimedia and Expo (ICME) | 979-8-3315-9495-4/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICME59968.2025.11208942

***Abstract*—In modern healthcare, emotional well-being is critical to patient recovery and overall outcomes. However, limited availability of trained professionals and time constraints often hinder the delivery of consistent emotional support. To address this gap, we propose the Embodied Emotional Care System (EECS), a comprehensive humanoid robotic framework designed to deliver personalized emotional care through an integrated, multi-layered architecture. EECS analyzes dynamic facial expressions and real-time vocal inputs to extract the patient's emotional state and semantic information, constructs context-aware prompts processed by an LLM for reasoning, and ultimately generates empathetic dialogues synchronized with human-like facial expressions and natural body movements to address diverse emotional support needs. Experimental results show that deploying EECS on a humanoid robot significantly boosts patient engagement through real-time multimodal interaction, delivering deeper emotional support and a more human-like therapeutic experience. Furthermore, it bridges gaps in professional emotional support resources, offering a feasible pathway to improve overall healthcare quality.**

***Index Terms*—emotional care system, dynamic facial expression recognition, large language model, motion generation**

I. Introduction

In recent years, there has been an increasing recognition of the critical role that emotional well-being plays in the efficacy of medical treatment and rehabilitation. Mental health care, supportive psychotherapy, and palliative interventions frequently demand sustained emotional engagement, empathy, and personal attention. However, the persistent shortage of trained mental health professionals and caretakers, coupled with high patient-to-clinician ratios, limits the extent and quality of individual emotional support that patients receive in healthcare environments. This challenge has stimulated research into innovative, technology-driven modalities capable of delivering personalized emotional care, with the ultimate aim of enhancing patient adherence, alleviating psychological distress, and improving clinical outcomes.

Against this backdrop, humanoid robots have emerged as promising platforms to address the gap in emotional care-giving. Compared with purely utilitarian or non-humanoid robotic forms, humanoid robots are inherently more intuitive and relatable, thus facilitating more natural and human-like interactions [1]. Their anthropomorphic design and articulated facial expressions can potentially reduce patient apprehension, foster trust, and encourage participation. Early attempts in robotics-assisted therapy, often involving social and companion robots, have shown some promise in providing emotional support in settings such as geriatric care, pediatric wards, and cognitive rehabilitation. Yet, current approaches often lack robust emotional intelligence, multi-modal sensing, and the capacity for nuanced, context-sensitive dialogue [2]. Systems frequently rely on single-modality emotional inputs (e.g., speech-only or facial expression-only), employ simplistic rule-based conversational models, or present static and unengaging robotic facial features. These shortcomings restrict the depth and authenticity of the emotional rapport that can be built and, consequently, limit the therapeutic utility of such robotic assistants.

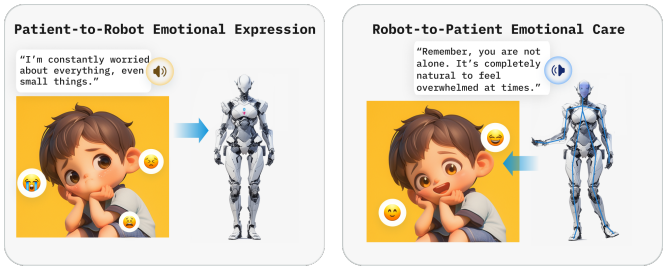

Fig. 1. Illustration of the bidirectional interaction in the Embodied Emotional Care System (EECS), where the humanoid robot equipped with EECS perceives the patient's emotions and semantic input, responding with human-like dynamic facial expressions, empathetic dialogue, and supportive motions to provide emotional care.

To address the lack of emotional support in therapeutic and medical scenarios, this study proposes an integrated Embodied Emotional Care System (EECS) designed specifically for humanoid robots, enabling emotion recognition, context-

*Corresponding author.

aware reasoning, and embodied emotional response generation. As illustrated in Figure 1, EECS employs Dynamic Facial Expression Recognition (DFER) and Automatic Speech Recognition (ASR) technologies to analyze facial expressions and vocal inputs in real time, extracting emotional states and semantic information. These multimodal signals are processed by a context-aware Large Language Model (LLM) to generate empathetic multi-task therapeutic dialogues. To enhance emotional expressiveness, EECS integrates a motion generation module and an audio-driven facial synthesis module, producing synchronized human-like body motions, and expressive facial cues. This framework, through multimodal coordinated embodied responses, transforms the robot into an "emotional agent," effectively improving patient engagement and delivering meaningful emotional support.

The contributions of this paper are as follows:

- **Proposed a Novel Framework for Embodied Emotional Care:** Developed the Embodied Emotional Care System (EECS), a multi-layered framework that seamlessly integrates multimodal recognition, context-aware reasoning, and embodied emotional response generation for humanoid robots, addressing critical needs in emotional care and support scenarios.
- **Designed Context-Aware Therapeutic Dialogue Prompts:** Designed a novel method that integrates emotional signals with contextual semantics, enabling the generation of empathetic dialogues, synchronized facial expressions, and body motions. This approach bridges emotional recognition with embodied systems, facilitating coherent and emotionally rich interactions in diverse scenarios;
- **Validated the Effectiveness of the EECS Framework:** Experimental evaluations across diverse emotional and age-based scenarios demonstrated the effectiveness of the EECS framework in enhancing emotional engagement and support, validated the critical role of emotional input in embodied emotional systems, highlighting the system's broad applicability in therapeutic care, emotional relief, and other emotion-driven support contexts.

II. RELATED WORK

Dynamic Facial Expression Recognition (DFER) and Automatic Speech Recognition (ASR): Dynamic Facial Expression Recognition (DFER) focuses on capturing and analyzing facial expression dynamics from sequential image data, such as videos. Current DFER approaches primarily rely on spatiotemporal feature extraction techniques to interpret the temporal evolution of expressions [3]. Transformer-based models, leveraging deep attention mechanisms, have emerged as a key technology due to their superior ability to handle complex spatiotemporal relationships [4]. Recent works like CLIPER and DFER-CLIP integrate the powerful prior knowledge of the vision-language model CLIP, fusing visual information with linguistic embeddings to enhance robustness and accuracy in expression recognition [5]. Additionally, Masked Autoencoders for Dynamic Facial Expression Recognition (MAE-DFER) utilize self-supervised learning to strengthen spatiotemporal modeling capabilities, providing new insights for handling complex dynamic relationships [6].

Automatic Speech Recognition (ASR) has undergone significant advancements, transitioning from Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) to deep neural network (DNN)-based acoustic modeling [7]. DNN methods have substantially improved recognition accuracy by refining acoustic representations. Recent systems, such as Alibaba's SenseVoice and OpenAI's Whisper, have achieved near-human-level performance through large-scale pretraining and Transformer-based architectures. These models leverage multimodal data to enable efficient decoding of speech content, providing a robust foundation for real-time emotional analysis and complex interactive scenarios.

Emotional Text Generation and Prompt Engineering: Emotional text generation has transitioned from traditional rule-based systems to transformer-based architectures capable of generating empathetic, contextually rich dialogues. State-of-the-art models like GPT-4, GPT-4o, Gemini, Llama3, and Qwen2 dominate the field, offering unparalleled language understanding and generation capabilities [8] [9] [10] [11]. Techniques such as fine-tuning with emotion-labeled datasets, incorporation of emotion tokens, and emotion-consistent decoding have enabled these models to align generated responses with desired emotional tones. Prompt engineering has emerged as a critical tool, with approaches like Few-Shot Prompting, Chain-of-Thought (CoT) prompting, and Self-Consistency Decoding enhancing the adaptability of large language models (LLMs) to emotional contexts. Advances in dynamic prompt generation, including retrieval-augmented techniques and structured templates, enable context-aware dialogue systems to manage multiple tasks with emotional depth.

Emotional Speech, Facial, and Motion Generation for Humanoid Robots: The generation of emotional outputs in humanoid robots encompasses speech synthesis, facial expression generation, and motion modeling. Emotional speech synthesis has seen significant advancements with models like Tacotron 2, FastSpeech 2, Flowtron, and VITS, which integrate emotion embeddings and prosody control for high-quality, expressive speech [12]. Lip and facial movement synchronization, achieved through models like LipGAN, Wav2Lip, and Speech2Face, ensures temporal alignment of speech with facial expressions [13]. StyleGAN2 and DiscoFaceGAN have set benchmarks in generating expressive facial features with high fidelity and fine-grained control [14]. Motion generation, critical for embodied interactions, has been enhanced through Transformer-based models such as MotionBERT and hierarchical reinforcement learning approaches like Action2Motion [15]. These models generate natural and emotion-consistent gestures and full-body motions by incorporating multimodal signals from speech and text. Physics-informed frameworks integrating kinematic and dynamic constraints, combined with reinforcement learning, have improved motion fluidity and realism. Multimodal frameworks that syn-

chronize emotional speech, facial expressions, and motion offer promising avenues for creating cohesive and engaging humanoid robots. Despite these advances, bridging multimodal emotional signals in humanoid robots to achieve a fully integrated, context-aware emotional care system remains an ongoing challenge.

III. METHODOLOGY

A. System Overview

The Embodied Emotional Care System (EECS) is a comprehensive, multi-layered framework that enables humanoid robots to deliver empathetic, multimodal emotional care by processing visual and auditory signals in real-time. Each layer is designed with specialized models and algorithms to address distinct aspects of emotional recognition, reasoning, and response synthesis. The system architecture, illustrated in Figure 2, is composed of five sequential layers:

- **Physical Layer:** Captures raw sensory data using cameras and microphones, representing visual and auditory inputs respectively.
- **Perception Layer:** Encodes the raw input signals into feature-rich representations suitable for emotion and semantic analysis.
- **Recognition Layer:** Performs multimodal emotional and semantic recognition, extracting emotional states and transcribed speech data.
- **Reasoning Layer:** Utilizes a Large Language Model (LLM) with context-aware prompt engineering to generate empathetic and task-oriented responses.
- **Generation Layer:** Synthesizes multimodal outputs, including speech, dynamic facial expressions, and motion, to convey human-like emotional responses.

This pipeline ensures seamless integration between input processing, emotional reasoning, and output generation. By leveraging state-of-the-art models and algorithms at each stage, EECS provides real-time, synchronized, and human-like interactions tailored to the emotional and therapeutic needs of patients.

B. Physical Layer

The Physical Layer serves as the interface between the external environment and the system, capturing multimodal input data through cameras and microphones. This layer ensures that raw sensory signals are acquired with sufficient fidelity to support downstream processing tasks. Specifically, it captures two primary types of input:

- **Facial Expression Video Stream:** Visual input is captured as a sequence of image frames from a camera, denoted as:

$$D_v = \{I_t \mid I_t \in \mathbb{R}^{H \times W \times C}, t = 1, \ldots, T\} \tag{1}$$

  where I_t represents a single video frame, H and W are the frame height and width, C is the number of color channels, and T is the total number of frames.

- **Audio Stream:** Auditory input is captured as sampled waveforms from a microphone, represented as:

$$D_a = \{A_t \mid A_t \in \mathbb{R}^S, t = 1, \ldots, T'\} \tag{2}$$

  where A_t is the sampled audio waveform, S is the sampling rate, and T' is the number of temporal segments in the audio signal.

C. Perception Layer

The Perception Layer processes raw input data from the Physical Layer, transforming visual and auditory signals into compact, feature-rich representations that are suitable for downstream emotional and semantic analysis. This layer employs advanced encoding models to extract high-dimensional embeddings from both video and audio streams.

- **Video Encoder:** The facial expression video stream D_v is processed by a convolutional encoder f_{encode}, which extracts spatiotemporal features from the sequence of frames:

$$F_{\text{face}} = f_{\text{encode}}(D_v), \quad F_{\text{face}} \in \mathbb{R}^{T \times d_v} \tag{3}$$

  where F_{face} represents the feature matrix, T is the number of frames, and d_v is the dimensionality of the extracted features.
- **Audio Encoder:** The audio stream D_a is processed by an audio encoder f_{encode} to generate a feature representation:

$$F_{\text{audio}} = f_{\text{encode}}(D_a), \quad F_{\text{audio}} \in \mathbb{R}^{T' \times d_a} \tag{4}$$

  where F_{audio} is the audio feature matrix, T' is the number of temporal audio segments, and d_a is the feature dimensionality.

D. Recognition Layer

The Recognition Layer is responsible for analyzing the encoded features from the Perception Layer to extract emotional states and semantic information. It employs advanced models for both Dynamic Facial Expression Recognition (DFER) and Automatic Speech Recognition (ASR) processing, enabling a comprehensive understanding of the patient's multimodal inputs.

- **Dynamic Facial Expression Recognition (DFER):** Temporal dependencies in the facial expression features F_{face} are analyzed using a DFER module g_{DFER}, which classifies the patient's emotional state:

$$Emotion = g_{\text{DFER}}(F_{\text{face}}), \quad Emotion \in \mathbb{R}^k \tag{5}$$

  where E_{emotion} is the extracted emotional vector, and k represents the number of predefined emotion classes (e.g., anger, fear, sadness, surprise).
- **Automatic Speech Recognition (ASR):** The audio feature matrix F_{audio} is processed by an ASR module h_{ASR}, which converts audio signals into text:

$$Text = h_{\text{ASR}}(F_{\text{audio}}) \tag{6}$$

  where $Text$ represents the transcribed textual content of the patient's speech.

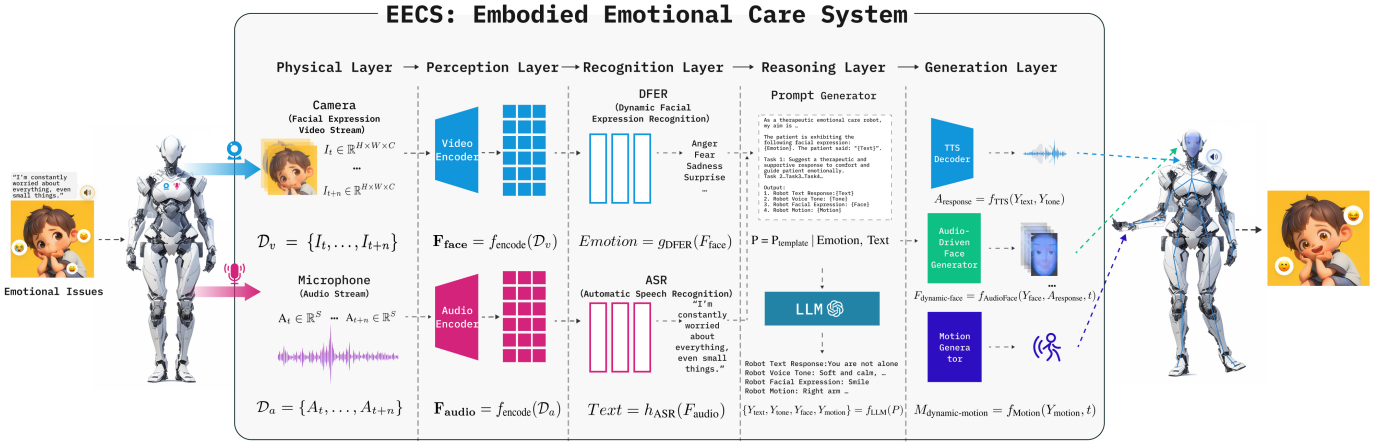

Fig. 2. Architecture of the Embodied Emotional Care System (EECS), illustrating the layered structure: the Physical Layer captures multimodal inputs through cameras and microphones; the Perception Layer processes video and audio signals via encoders; the Recognition Layer performs Dynamic Facial Expression Recognition (DFER) and Automatic Speech Recognition (ASR) to extract emotions and text; the Reasoning Layer generates context-aware prompts processed by a large language model (LLM); and the Generation Layer synthesizes empathetic responses, including audio, dynamic facial expressions, and motion, to provide comprehensive emotional care.

E. Reasoning Layer

The Reasoning Layer employs a context-aware prompt generation mechanism and a LLM to produce empathetic, task-specific responses. It combines emotional cues and semantic content extracted from the Recognition Layer to construct a structured prompt. The LLM processes this prompt to generate multi-task outputs (text, tone, facial expression, motion) tailored to the patient's therapeutic needs.

Prompt Design for Emotional Care Robot

Objective: Understand the patient's emotional state, provide empathetic and supportive responses, and use appropriate face expressions, tone, and motion to aid in emotional recovery.

Input: Facial expression: {Emotion}.Patient statement: "{Text}". **Tasks:** Suggest a supportive response to guide the patient emotionally; Provide the emotional tone for voice output; Specify a suitable facial expression for the robot; Suggest a robot motion to reinforce emotional support. **Output:** Robot Text Response: {Text}; Robot Voice Tone: {Tone}; Robot Facial Expression: {Face}; Robot Motion: {Motion}

The prompt P is generated by combining the template P_{template}, the emotional state $Emotion$, and the transcribed text $Text$:

$$P = P_{\text{template}} \,|\, \text{Emotion, Text} \tag{7}$$

The LLM processes P to generate outputs:

$$\{Y_{\text{text}}, Y_{\text{tone}}, Y_{\text{face}}, Y_{\text{motion}}\} = f_{\text{LLM}}(P) \tag{8}$$

where:

- Y_{text}: Textual response.
- Y_{tone}: Voice tone suggestion.
- Y_{face}: Facial expression parameters.
- Y_{motion}: Motion parameters.

F. Generation Layer

The Generation Layer is responsible for synthesizing multimodal outputs, including speech, dynamic facial expressions, and motion, to provide a cohesive and empathetic response. This layer takes the outputs generated by the Reasoning Layer and processes them through dedicated modules for Text-to-Speech (TTS), Audio-Driven Facial Expression Generation, and Motion Generation.

1. Text-to-Speech (TTS): The textual response Y_{text} and the emotional tone Y_{tone} are processed by the TTS decoder f_{TTS}, which generates the audio response A_{response}:

$$A_{\text{response}} = f_{\text{TTS}}(Y_{\text{c}}, Y_{\text{tone}}) \tag{9}$$

where A_{response} represents the synthesized speech output.

2. Audio-Driven Facial Expression Generation: The facial expression parameters Y_{face} and the audio response A_{response} are processed by the facial expression generator $f_{\text{AudioFace}}$. This module generates time-synchronized dynamic facial expressions $F_{\text{dynamic-face}}$:

$$F_{\text{dynamic-face}} = f_{\text{AudioFace}}(Y_{\text{face}}, A_{\text{response}}, t), \quad t \in [1, n] \tag{10}$$

where t represents the time frame, and n is the total number of time steps.

3. Motion Generation: The body movement parameters Y_{motion} are processed by the motion generator f_{Motion} to produce dynamic body movements $M_{\text{dynamic-motion}}$:

$$M_{\text{dynamic-motion}} = f_{\text{Motion}}(Y_{\text{motion}}, t), \quad t \in [1, n] \tag{11}$$

where $M_{\text{dynamic-motion}}$ represents motion outputs.

The outputs A_{response}, $F_{\text{dynamic-face}}$, and $M_{\text{dynamic-motion}}$ are synchronized and delivered in real-time to ensure a seamless and emotionally coherent interaction. This layer ensures that

the robot's multimodal responses effectively convey empathy and emotional support.

IV. EXPERIMENTS

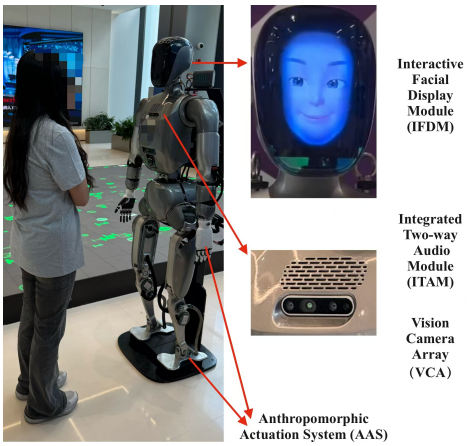

Fig. 3. The GuangHua Humanoid Robot integrates core modules for multimodal interaction: Interactive Facial Display Module (IFDM), Integrated Two-way Audio Module (ITAM), Vision Camera Array (VCA), and Anthropomorphic Actuation System (AAS).

A. System Deployment and Technical Implementation

To validate the effectiveness of the Embodied Emotional Care System (EECS), the experiment implemented EECS using advanced models and hardware to achieve real-time multimodal emotional interaction. Automatic Speech Recognition (ASR) utilized the OpenAI Whisper model (809M parameters) for high-accuracy transcription, while emotional recognition employed the MAE-DFER model (84.9M parameters) for robust dynamic facial expression analysis [6]. The system integrated Qwen2.5 (14B parameters) as the Large Language Model (LLM) to generate context-aware and empathetic dialogues [16]. Emotional Text-to-Speech (TTS) was powered by EmotiVoice, real-time facial expression and lip synchronization were managed by MuseV and MuseTalk, and motion responses were executed using a kinematic motion control system to ensure precise and emotion-aligned motions [17]. The system operated on one NVIDIA A6000 GPU (48GB VRAM), enabling efficient processing of computationally intensive multimodal interactions required for real-time therapeutic applications.

As shown in Figure 3, the entire EECS system was deployed on the GuangHua humanoid robot. This platform integrates key components, including an Interactive Facial Display Module (IFDM) for visualizing dynamic emotional expressions, an Integrated Two-way Audio Module (ITAM) for high-fidelity audio input and playback, a Vision Camera Array (VCA) for real-time facial expression recognition, and an Anthropomorphic Actuation System (AAS) for executing coordinated, emotion-aligned gestures and full-body motions [18].

B. Experimental Design and Evaluation Criteria

To evaluate the effectiveness of the Embodied Emotional Care System (EECS), an experimental design was implemented involving 100 participants to ensure diversity across age and gender groups. The primary goal of this experiment was to assess the system's ability to deliver emotional support and provide a meaningful care experience across a variety of emotional contexts. The experiment focused on the core metric of "Care Experience," measured on a 5-point Likert scale.

Participants: The experiment included 100 participants with an equal gender distribution (50% male and 50% female). To capture the impact of age on care experience, participants were divided into three age groups:

- *Youth group (14–24 years):* Representing individuals in early emotional development stages.
- *Adult group (25–40 years):* Representing individuals with mature emotional and social capabilities.
- *Senior group (41–60 years):* Representing individuals with diverse and complex emotional needs.

Experimental Procedure: Each participant was asked to role-play five distinct emotional states, representing common negative emotional challenges—*Anxiety, Sadness, Anger, Fear, and Frustration*. In each role-played scenario, EECS engaged with participants using multimodal interactions, including speech-based dialogue, real-time facial expressions and motions. Participants then evaluated how effectively the system provided supportive responses aligned with their simulated emotional state.

Evaluation Metric: After each interaction, participants were asked to rate their "Care Experience" on a 5-point Likert scale (1 = Poor, 2 = Fair, 3 = Neutral, 4 = Good, 5 = Excellent). The scale was designed to capture user satisfaction with the system's emotional support.

C. Experimental Results and Performance Analysis

TABLE I
CARE EXPERIENCE SCORES BY EMOTION TYPE, GENDER, AGE GROUP

Dimension	Category	Mean Score
Emotion Type	Anxiety	4.7
	Sadness	4.6
	Anger	4.4
	Fear	4.4
	Frustration	4.6
	Total	4.54
Gender	Male	4.5
	Female	4.6
	Total	4.55
Age Group	14–24 (Youth)	4.6
	25–40 (Adult)	4.4
	41–60 (Senior)	4.5
	Total	4.5

The experimental results were analyzed independently based on Emotion Type, Gender, and Age Group to evaluate the performance of the Embodied Emotional Care System (EECS).

TABLE II
COMPARISON OF EECS WITH AND WITHOUT EMOTIONAL INPUT

Patient's Emotion	Patient's Input	Robot's Response	Robot's Tone	Robot's Expression	Robot's Motion	Score
Anxiety	"I feel overwhelmed and don't know what to do."	"You're not alone, I can help you step by step."	Encouraging	Smile	Leaning forward, nodding	5
	"I feel overwhelmed and don't know what to do."	"What's the issue? Can you explain more?"	Neutral	Neutral expression	No motion	3
Anger	"Why is everything so unfair to me?"	"I understand you. Let's address this together calmly."	Gentle	Warm gaze	Arms open for a hug	5
	"Why is everything so unfair to me?"	"Why do you think so? Let's discuss."	Neutral	Neutral expression	No motion	2

As shown in Table I, the overall Care Experience scores are consistently between "good" and "excellent," demonstrating the outstanding effectiveness of EECS in delivering emotional support and validating its value in multimodal interaction scenarios.

As indicated in Table I, the independent analysis reveals key insights. Intense emotions such as Anger and Fear exhibit slightly lower scores, suggesting that calming such emotions might be more challenging. Female participants provided higher ratings compared to males, indicating a higher affinity for the system. Additionally, both younger (14–24 years) and older (41–60 years) participants showed greater satisfaction compared to the middle age group (25–40 years), suggesting age-related variations in receptiveness to robotic care.

We conducted an additional experiment to evaluate the performance of the EECS system with and without emotional input. As shown in Table II, the system with emotional input produced responses that were more encouraging and empathetic, accompanied by expressive facial cues and meaningful gestures, such as leaning forward or opening arms for a hug, resulting in significantly higher scores. In contrast, without emotional input, the system exhibited neutral responses, minimal expressions, and a lack of supportive motions, leading to lower scores. This comparison underscores the critical role of emotional input in enhancing the effectiveness of the EECS system and demonstrates its importance in delivering more human-like and emotionally supportive interactions.

V. CONCLUSION

This study proposed the Embodied Emotional Care System (EECS), a multi-layered framework designed to enable humanoid robots to deliver comprehensive emotional support through multimodal interaction. EECS integrates advanced dynamic facial expression recognition, automatic speech recognition, context-aware reasoning based on a large language model, and synchronized emotional multimodal output generation, demonstrating robust performance in emotion recognition and emotional care interactions. Experimental results show that EECS performs exceptionally well in multi-emotion scenarios, achieving high user satisfaction and validating its effectiveness in psychological support and emotional caregiving. The system's innovative design addresses a critical gap in embodied emotional support, and future research will focus on enhancing its adaptability and expanding its real-world applications.

REFERENCES

[1] Jingwen Hou, Weisi Lin, Guanghui Yue, Weide Liu, and Baoquan Zhao, "Interaction-matrix based personalized image aesthetics assessment," *IEEE Transactions on Multimedia*, vol. 25, pp. 5263–5278, 2022.

[2] Jiebin Yan, Lei Wu, Yuming Fang, Xuelin Liu, et al., "Video quality assessment for online processing: From spatial to temporal sampling," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[3] Weide Liu, Xiangfei Kong, et al., "Cross-image region mining with region prototypical network for weakly supervised segmentation," *IEEE Transactions on Multimedia*, vol. 25, pp. 1148–1160, 2021.

[4] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, et al., "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, pp. 1–74, 2021.

[5] Hanting Li, Hongjing Niu, et al., "Cliper: A unified vision-language framework for in-the-wild facial expression recognition," in *2024 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2024, pp. 1–6.

[6] Licai Sun, Zheng Lian, et al., "Mae-dfer: Efficient masked autoencoder for self-supervised dynamic facial expression recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6110–6121.

[7] Sparsh Mittal, "A survey on modeling and improving reliability of dnn algorithms and accelerators," *Journal of Systems Architecture*, vol. 104, pp. 101689, 2020.

[8] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[9] Gemini Team, Rohan Anil, et al., "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[10] Abhimanyu Dubey, Abhinav Jauhri, et al., "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[11] Jinze Bai, Shuai Bai, et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[12] Yuxuan Wang, RJ Skerry-Ryan, et al., "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[13] KR Prajwal, Rudrabha Mukhopadhyay, et al., "A lip sync expert is all you need for speech to lip generation in the wild," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 484–492.

[14] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny, "Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3626–3636.

[15] Wentao Zhu, Xiaoxuan Ma, et al., "Motionbert: A unified perspective on learning human motion representations," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15085–15099.

[16] An Yang, Baosong Yang, et al., "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[17] Yue Zhang, Minhao Liu, et al., "Musetalk: Real-time high quality lip synchronization with latent space inpainting," *arXiv preprint arXiv:2410.10122*, 2024.

[18] Zhonghan Lin, Qi Shao, Xin-Jun Liu, and Huichan Zhao, "An anthropomorphic musculoskeletal system with soft joint and multifilament pneumatic artificial muscles," *Advanced Intelligent Systems*, vol. 4, no. 10, pp. 2200126, 2022.