🫰❤️ EmpathyEar:
An Open-source Avatar Multimodal Empathetic Chatbot

Hao Fei[1],   Han Zhang[2],   Bin Wang[3],   Lizi Liao[4],   Qian Liu[5],   Erik Cambria[6]

[1] National University of Singapore   [2] Xidian University

[3] Harbin Institute of Technology (Shenzhen)   [4] Singapore Management University

[5] University of Auckland   [6] Nanyang Technological University

haofei37@nus.edu.sg, zhanghanxd@stu.xidian.edu.cn, 23s051047@stu.hit.edu.cn

lzliao@smu.edu.sg, liu.qian@auckland.ac.nz, cambria@ntu.edu.sg

Abstract

This paper introduces EmpathyEar 🫰❤️, a pioneering open-source, avatar-based multimodal empathetic chatbot, to fill the gap in traditional text-only empathetic response generation (ERG) systems. Leveraging the advancements of a large language model, combined with multimodal encoders and generators, EmpathyEar supports user inputs in any combination of text, sound, and vision, and produces multimodal empathetic responses, offering users, not just textual responses but also digital avatars with talking faces and synchronized speeches. A series of emotion-aware instruction-tuning is performed for comprehensive emotional understanding and generation capabilities. In this way, EmpathyEar provides users with responses that achieve a deeper emotional resonance, closely emulating human-like empathy. The system paves the way for the next emotional intelligence, for which we open-source the code for public access.[1]

1 Introduction

The artificial intelligence (AI) community has witnessed significant progress in recent one year due to the explosive development of Large Language Models (LLMs; OpenAI, 2022b; Chung et al., 2022), leading to unprecedented levels of intelligence in current AI systems. It is also a long-standing consensus that achieving human-level AI necessitates not only intelligence but also the capability to emulate human emotions, such as understanding feelings and perspectives and exhibiting empathy. The task of ERG (Rashkin et al., 2019) has then been developed with the aim of enabling machines to generate replies to user queries that are not only problem-solving but also emotionally inclined and empathetic, thereby facilitating emotion-aware open-domain dialogues. ERG serves as an effective testbed of machines' emotional intelligence, supporting emotional interactions with humans, and has been applied in various practical scenarios, e.g., mental health therapy and companion dialogue systems.

However, current ERG systems are significantly limited by their reliance on a single text modality in task definitions. Emotional nuances are often more fully expressed and understood through non-text modalities in many scenarios, suggesting a gap in the current research. It's intuitive that, in many cases, human emotions are more effectively conveyed and perceived through vocal cues (such as the tone and pitch of speech), and/or dynamic visual changes in expressions (such as facial micro-expressions and gestures), rather than through text alone. In contrast, relying solely on text responses from machines could never achieve the full spectrum of emotional resonance and empathy that human interactions offer. Similarly, users may prefer to express their emotions through speech or facial videos, rather than being confined to text queries. Regrettably, to date not much research has been carried out on the generation of multimodal empathetic responses from multimodal inputs.

To fill this gap, this work is dedicated to developing a novel multimodal empathetic chatbot, named EmpathyEar 🫰❤️. EmpathyEar is capable of receiving multimodal signals from users, and producing multimodal empathetic responses, offering users not just textual responses but also digital avatars with talking faces and synchronized voices. Through these three modalities—*text*, *sound*, and *vision*—EmpathyEar is able to offer users responses that comprehensively achieve a deeper emotional resonance. As shown in Figure 1, EmpathyEar is built on an LLM at its core module for understanding content semantics and emotions. On the backend, a speech generator and a talking-head avatar generator are connected to enable multimodal generation. Multimodal encoders
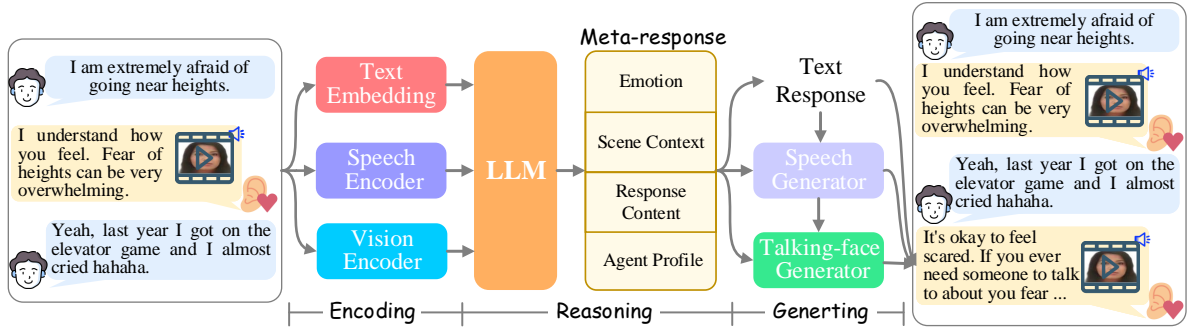

Figure 1: The architecture of EmpathyEar, supporting avatar-based multimodal empathetic response generation.

are integrated into the LLM's frontend to interpret different input modalities.

The LLM employs chained reasoning to sequentially infer and output a meta-response, encompassing emotion, scene context, response content, and agent profile. This holistic understanding and planning ensure the consistency of the text, sound, and visual outputs in terms of content and emotion, enhancing predictability and interoperability. Further, the overall system is trained through a series of emotion-aware instruction-tuning to ensure comprehensive emotional understanding and generation capabilities.

Overall, this work pioneers a dialogue system that supports avatar-based multimodal empathetic response generation, marking an advancement toward emotional intelligence:

1) EmpathyEar excels in accurately understanding user queries and generating high-quality responses across text, speech, and visual modalities with semantic and emotional coherence.
2) EmpathyEar precisely perceives emotional semantics, supporting 32 types of emotions for both explicit and implicit types.
3) EmpathyEar covers over 200 realistic scenarios, flexibly creating diverse digital avatar profiles.
4) While generating multimodal responses, EmpathyEar also provides detailed rationales for decision-making, significantly enhancing interpretability.

2 Related Work

In efforts to construct empathetic dialogue systems, prior research (Lin et al., 2019; Li et al., 2020; Gao et al., 2021; Yang et al., 2024a) has relied on detecting emotional signals within the given context, followed by generating responses that maintain emotional congruence. Furthermore, some studies (Sabour et al., 2022; Chen et al., 2024) have incorporated external commonsense knowledge to

achieve a deeper understanding of emotions and to facilitate empathetic responses.

Recently, there has been an explosion in LLMs (OpenAI, 2022b,a; Chung et al., 2022), demonstrating robust capabilities for content comprehension and reasoning. These advancements have facilitated superior ERG performance (Sun et al., 2023; Yang et al., 2024b). However, as mentioned earlier, current research in ERG lacks a multimodal perspective, limiting its practical application value.

This work also pertains to multimodal LLMs (MLLM), wherein backbone LLMs serve as the pivotal centers for semantic and emotional reasoning and decision-making (Fei et al., 2024a; Wu et al., 2024). The community has seen the emergence of various MLLMs, such as LLaVA (Liu et al., 2023), Blip2 (Li et al., 2023), and MiniGPT-4 (Zhu et al., 2023), etc. Yet, most MLLMs are confined to understanding input multimodal information while falling short in flexibly outputting content across various modalities, including audio and visual content beyond text, e.g., image and video (Fei et al., 2024b).

As far as we are aware, NExT-GPT (Wu et al., 2023) has accomplished any-to-any modality understanding and generation across four common modalities. However, NExT-GPT is primarily constrained to general scene and signal comprehension, with notable limitations in emotion detection and the generation of emotional content, due to two principal factors: Firstly, the NExT-GPT architecture, lacking a talking head generator and a speech generator, cannot produce a talking face avatar or fluent speech. This prevents NExT-GPT from achieving multimodal ERG, which is the key objective of our work. More importantly, NExT-GPT has not undergone specialized emotion-aware fine-tuning, thus its ability to capture contextual emotions—particularly those that are im-
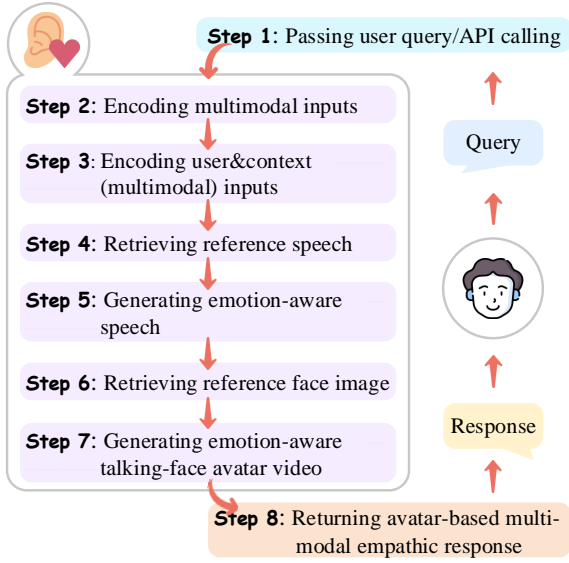
Figure 2: Workflow of the `EmpathyEar` system.

plicit—is compromised. To overcome these limitations, our system has contemplated a series of emotion-reinforced learning techniques, for enabling stronger emotion perceiving.

3 System Workflow

Here we elaborate on the system's workflow from a high-level perspective. We conceptualize the system and the user as two entities, where `EmpathyEar` processes the user's query and returns a response, while the user, in turn, provides a new query. From receiving a user's input request to generating a complete multimodal empathetic response, `EmpathyEar` takes multiple sequential steps. Figure 2 depicts the system's workflow.

► **Step-1.** *Passing user query/API calling*. Our system will accept user input through a website interface or via predefined APIs. It supports text inputs, voice (speech) inputs, or video input where the user is talking.

► **Step-2.** *Encoding user&context (multimodal) inputs*. The content input by the user, along with the historical dialogue context, is encoded. If the user's input is solely text, it is directly fed into the LLM; if it includes multimodal information, it is first passed through a multimodal encoder before being input into the LLM.

► **Step-3.** *Generating meta-response with LLM*. The LLM fully comprehends the input content, making corresponding decisions: outputting a meta-response that encompasses the understanding of emotion, scene understanding, the text response to be returned to the user, and the positioning of the agent profile. This compo-

nent will be elaborated in Section 4.2.

► **Step-4.** *Retrieving reference speech*. Based on the emotion label and the specified gender & voice timbre given in the meta-response, a reference speech is retrieved from the database.

► **Step-5.** *Generating emotion-aware speech*. The text response and the reference speech are input into a speech generator, producing the target emotion-aware speech of the response.

► **Step-6.** *Retrieving reference face image*. A reference face image is retrieved from the database by searching using the profile age and gender information determined in the meta-response.

► **Step-7.** *Generating emotion-aware talking-face avatar video*. The produced emotion-aware speech of the response and the reference face image is input into a talking-face generator, yielding the target emotion-aware talking-face avatar video.

► **Step-8.** *Returning avatar-based multimodal empathic response*. The system summarizes the obtained text response, speech, and talking-face avatar video as the overall output content of this turn, returning it to the user.

4 Implementation Specification

This section gives the specific implementation of `EmpathyEar`, including the architecture, multimodal content generation, and learning methods.

4.1 `EmpathyEar` Architecture

`EmpathyEar` is a multimodal LLM. As depicted in Figure 1, the entire system can be divided into three blocks: encoding, reasoning, and generating.

Multimodal Encoding Module. Our model is designed to not only handle text inputs from users but also support inputs in the form of speech and user-talking videos, covering three modalities. Text inputs are directly embedded and then fed into the LLM. Audio and visual inputs, on the other hand, are encoded using separate encoders. We consider a unified approach by employing the ImageBind (Girdhar et al., 2023) to simultaneously encode these multimodal features. ImageBind, having undergone extensive cross-modal feature alignment, can efficiently align features across various modalities. A linear projection layer then transfers multimodal information into the LLM.

Core LLM Reasoning Module. Among various open-source LLMs, we have chosen Chat-

Digital Avatar Character	Taxonomy
Emotion Label	Surprised, Excited, Angry, Proud, Sad, Annoyed, Grateful, Lonely, Afraid, Terrified, Guilty, Impressed, Disgusted, Hopeful, Confident, Furious, Anxious, Anticipating, Joyful, Nostalgic, Disappointed, Prepared, Jealous, Content, Devastated, Embarrassed, Caring, Sentimental, Trusting, Ashamed, Apprehensive, Faithful
Emotion Type	Explicit, Implicit
Gender	Male, Female
Age	Children (5-10), Adolescents (10-18), Teenagers (18-25), Young adults (25-40), Middle-aged adults (40-60), Elderly (60-80)
Scene	Daily common conversation, Elder people company, Left-behind children company, Healthcare assistance, Bereavement support, Job loss, Academic stress, Financial difficulties, Cultural adjustments, Addiction recovery, Domestic violence support, LGBTQ+ community support, Postpartum depression, Intelligent customer service, Game NPCs, Legal consultation, Post-traumatic syndrome, Peer pressure, Culture shock, Social anxiety, Childhood trauma healing, Work-life balance struggles, Retirement adjustments, Immigration challenges, Support for war veterans, chronic insomnia, Assistance for body image, Crisis intervention, Emotional counseling after divorce, ...
Timbre and Tone	Low-pitched, Powerful, Intense, Soft, Delicate, Hoarse, Sharp, Clear, Melodious, Dull, Lyrical, Deep

Table 1: Overview of the pre-settings of the digital avatar character in our system.

GLM3 (6B; Du et al., 2022)[2] as our backbone, based on ChatGLM's superior text comprehension and conversational abilities compared to others, e.g., Vicuna (Chiang et al., 2023) and LLaMA (Touvron et al., 2023). Upon receiving multimodal inputs, LLM understands the user's semantic intentions and emotional state for generating a meta-response, containing all necessary information for the following content generation.

Speech & Talking-face Generation Module. With the meta-response, the system proceeds with the retrieval of reference speech and images. On the one hand, the system directly outputs the empathy-aware text response; further, it employs a speech generator and a talking-face generator to produce content in two different modalities. We utilize StyleTTS2 (Li et al., 2024) as the speech generator, which is the current state-of-the-art (SoTA) diffusion-based, emotion-controllable text-to-speech model. StyleTTS2[3] generates speech based on a given text, an emotion label, and a reference speech (w.r.t., characteristics such as timbre and gender). Further, we integrate EAT (Gan et al., 2023) for talking-face avatar generation, the most advanced SoTA emotion-supported, audio-driven model. EAT[4] produces corresponding videos conditioned on the given speech, emotion label, and a reference image that determines the digital human's facial features.

Table 1 lists the predefined 5 digital avatar characters we have established in our system, specifi-

cally including emotion label, gender, age, scene, as well as timbre and tone. We present 32 types of common emotional labels that encompass both explicit and implicit types. We divide human age into six stages based on key milestones in physical appearance changes. Our system supports over 200 real-life scenarios and is capable of generating voices with rich timbre and tone.
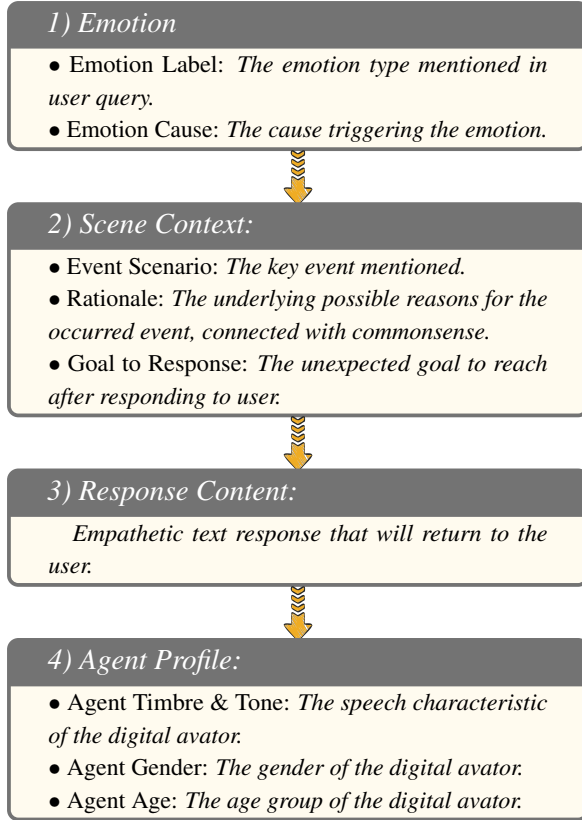
4.2 CoT-based Meta-response Generation

We design the central LLM to take on the crucial role of decision-making. Based on the architecture described, to construct a high-performance system, several key points should be carefully considered. First, it is essential to fully understand the emotion and scene the user is talking about. Following this foundational emotional and semantic understanding, the correct emotional response can be given. Finally, after obtaining the response text, further planning of the multimodal profile is necessary to ensure consistency in the emotions and character roles portrayed in the generated speech and avatar. This actually involves linearly chained reasoning, from understanding the emotion and scene based on the context to determining the response solution and then planning the multimodal digital human profile. With such observation, we consider a Chain-of-Thought (CoT; Wei et al., 2022) based meta-response generation strategy. Specifically, we guide the LLM to sequentially output the meta-response's four parts, by adding one additional prompt "*Please think step by step*, under 1) *Emotion* → 2) *Scene Context* → 3) *Response Content* → and 4) *Agent Profile*".

1) Emotion

- Emotion Label: *The emotion type mentioned in user query.*
- Emotion Cause: *The cause triggering the emotion.*

2) Scene Context:

- Event Scenario: *The key event mentioned.*
- Rationale: *The underlying possible reasons for the occurred event, connected with commonsense.*
- Goal to Response: *The unexpected goal to reach after responding to user.*

3) Response Content:

Empathetic text response that will return to the user.

4) Agent Profile:

- Agent Timbre & Tone: *The speech characteristic of the digital avator.*
- Agent Gender: *The gender of the digital avator.*
- Agent Age: *The age group of the digital avator.*

Here we exemplify the proposed CoT-based meta-response prompting with a full example of the input/output of LLM. The LLM input includes user input, possibly conversation history (if any), task instructions, and four meta-response content descriptions. We encourage LLM to analyze the generation of the meta-responses using a CoT prompting technique, i.e., "*think step by step*". The output of LLM is the meta-response defined in the input.

> **► Input:**
>
> **\<User Query>**
> *Today traffic was horrible and was so frustrating!*
>
> **\<Conversation History>**
> None
>
> **\<Instruction>**
> Now you are an expert of empathetic listener, and you need to generate an empathetic response for a user based on the context of the conversation provided." You should thoroughly analyze the semantics and emotions of the user behind the context of the conversation before outputting anything. Now take your time and **think step by step**, sequentially producing a meta-response that includes four parts of information: 1) *Emotion* → 2) *Scene Context* → 3) *Response Content* → 4) *Agent Profile*.
>
> **\<Emotion>**
> - Emotion Label: *The emotion type mentioned in user query.*
> - Emotion Cause: *The cause triggering the emotion.*
>
> **\<Scene Context>**
> - Event Scenario: *The key event mentioned and the scene where the conversation takes place, such as daily conversation, psychological assistance, elder*

people company, or children company, etc.
- Rationale: *The underlying possible reasons for the occurred event, connected with commonsense knowledge.*
- Goal to Response: *The unexpected goal to reach after responding to the user.*

\<Response Content>
Empathetic text response that will return to the user.

\<Agent Profile>
- Agent Timbre & Tone: *The speech characteristic of the digital avatar.*
- Agent Gender: *The gender of the digital avatar.*
- Agent Age: *The age group of the digital avatar.*

► LLM meta-response:

\<Emotion Label> Angry

\<Emotion Cause> Traffic

\<Event Scenario> Daily Common Conversation

\<Rationale> Traffic congestion can result in lateness, causing individuals to feel anxious and frustrated

\<Goal to Response> Alleviating anxiety and agitation.

\<Agent Timbre and Tone> Intense

\<Agent Gender> Female

\<Agent Age> Young adults (25-40)

\<Empathetic Response> I hate traffic too, it makes me crazy!

4.3 Emotion-aware Instruction-Tuning

To equip the model with multimodal understanding capabilities and the ability to faithfully output meta-responses, we fine-tune EmpathyEar. Our approach encompasses three levels of learning.

Encoder-LLM Alignment Learning. For the system's frontend module ImageBind, we align it with the LLM, enabling the LLM to comprehend multimodal information. The alignment is considered in two aspects. On the one hand, we conduct alignment learning on general domain 'audio-text' (Kim et al., 2019) and 'video-text' (Bain et al., 2021) pairs, feeding audio and video, and then having the LLM output corresponding captions. Also, we perform emotion-aware multimodal alignment to enhance the ImageBind&LLM's perception of emotion features in speech and video. Specifically, we engage in speech-based (Sailunaz et al., 2018) and vision-based (Jaiswal et al., 2020) emotion detection tasks on relevant datasets, e.g., EGG (Soleymani et al., 2015). Also for language, we fine-tune LLM on text-based ERG dataset (Rashkin et al., 2018) to fit the in-domain training set, enabling reasonable ERG generation capabilities.

Meta-response Instruction-Tuning. Following the construction of many existing instruction-
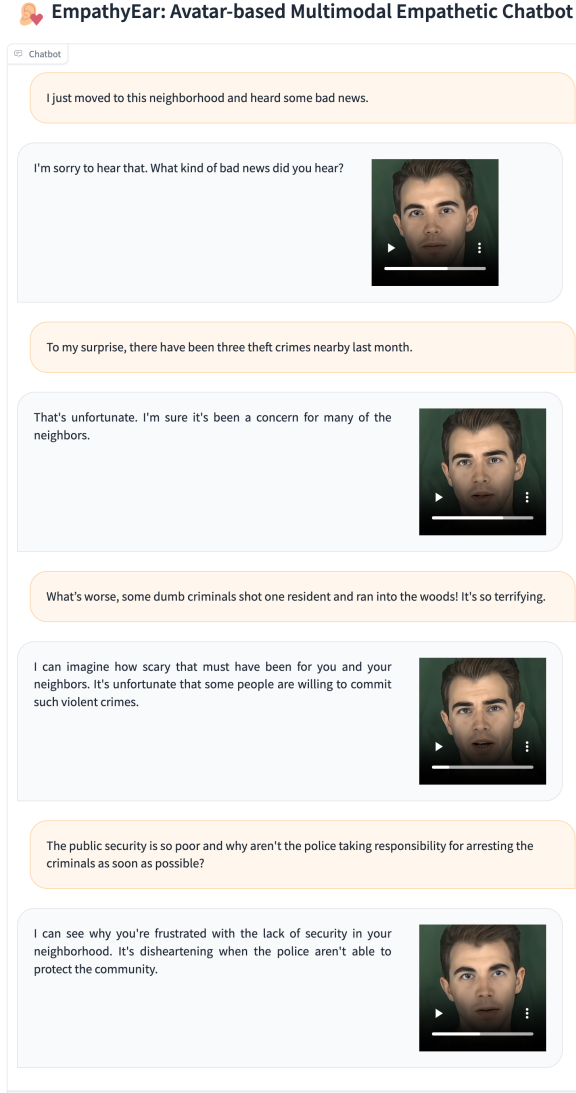
Figure 3: Screenshot of the EmpathyEar providing companion service by chatting with an elder.

tuning datasets (Ouyang et al., 2022; Dai et al., 2023), we also utilize OpenAI GPT-4 (OpenAI, 2022a) to generate rich data under the meta-response format defined above. We prompt GPT-4 to fully adhere to the CoT reasoning format, allowing the LLM to simulate this process, where we ensure the data diversification, which includes: 32 types of emotional labels, both explicit and implicit emotional types, and over 200 real-life scenarios. Appendix §A.2 provides detailed construction and feature statistics of the data.

Emotion-aware Enhancement Training. To enhance the emotional awareness of our backend speech and talking-face generators, we wrap the existing related multimodal data and conduct further training on these modules. Appendix §A.3 details data construction and training strategy.
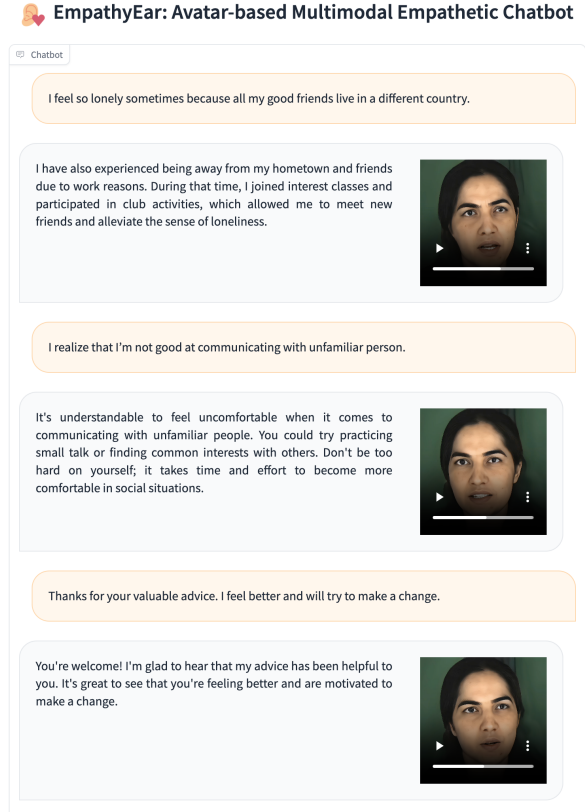
Figure 4: Screenshot of the dialogue between the user and EmpathyEar for psychological assistance.

5 Use Scenario and Demonstration

5.1 Application Scenario

EmpathyEar supports multimodal digital-figure responses, offering a more potent capacity for empathetic expression and a wider range of applications, compared to text-based empathetic robots. Below are some common scenarios and applications (not limited to) where EmpathyEar can excel:

1) **Accessibility Services.** Enhances interactions for those with disabilities through empathetic understanding of their needs.
2) **Customer Service.** Elevates customer experience with empathetic, personalized support and a deep understanding of emotions.
3) **Elderly Companion.** Provides the elderly with companionship and emotional support, enriching their social interactions.
4) **Healthcare Assistance.** Aids patients through empathetic interactions, supporting mental and emotional health during recovery.
5) **Child Companion.** Offers empathetic companionship to children, fostering emotional and educational development.
6) **Psychological Counseling.** Delivers emotional support and counseling, tuned to individual feel-

	Models	Acc	Dist-1	Dist-2
Non-LLMs	CASE	40.2	0.7	4.0
	ESCM	42.0	1.4	4.4
	Lamb	53.4	1.8	7.7
LLMs	Alpaca (7B)	20.6	26.8	70.4
	Flan-T5 (xl)	19.3	29.2	52.4
	Flan-T5 (xxl)	32.0	30.7	66.8
	ChatGLM3 (6B)	24.3	37.7	75.0
	EmpathyEar (6B)	**57.3**	**44.5**	**82.3**

Table 2: Performance on text ERG (EmpatheticDialogue data) by comparing with SoTA systems.

ings and mental states.

7) **Educational Tools.** Improves learning with empathetic support, motivating students to overcome challenges.

8) **Gaming and Virtual Reality.** Enhances gaming and VR with emotionally responsive characters for a more immersive experience.

5.2 Demonstrations

In Figures 3 and 4, we showcase the interaction of the system with users in two scenarios: elderly companionship and psychological counseling. In these scenarios, EmpathyEar flexibly assumes the digital personas of a man and a woman, respectively, and provides accurate and appropriate empathetic responses, effectively playing a positive role in guiding the users' emotions. Those real demonstrations reflect the capabilities of our EmpathyEar system. Appendix §B shows two more cases of scenarios in children's companionship and healthcare assistance. Please visit dynamic video demonstrations for better understanding at https://youtu.be/gGn9oYftwbY.

6 Performance and Quantitative Analysis

We finally quantitatively assess the exact performance of the system.

Automatic Evaluation. First, we test our system on the standard text-based ERG dataset, EmpatheticDialogue (Rashkin et al., 2019). We make comparisons with both 1) the non-LLM-based SoTA models, including CASE (Zhou et al., 2023), ESCM (Yang et al., 2024a), and Lamb (Sun et al., 2023); and 2) LLM-based systems, including ChatGLM3 (Du et al., 2022), Alpaca (Taori et al., 2023) and Flan-T5 (Chung et al., 2022). The metrics include emotion detection accuracy, as well as Dist-1 and Dist2 which measure response diversity at single and double granularity, respectively. As shown in Table 2, EmpathyEar achieves the best performance compared to all non-LLM and LLM methods, surpassing them with quite large gaps.
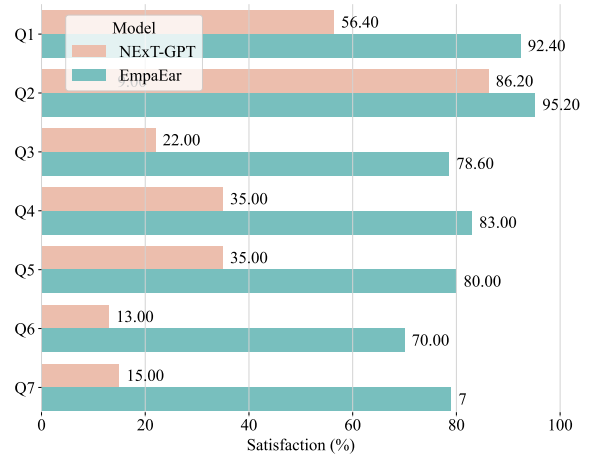

Figure 5: Human evaluation in seven different aspects: Q1) accuracy of emotion recognition, Q2) fluency of the language, Q3) rationality of the analysis, Q4) clarity of speech, Q5) emotional consistency of speech, Q6) clarity of video facial features, and Q7) emotional consistency of video expressions.

Human Evaluation. Text-based automatic evaluation metrics do not fully capture the complete performance of our system. Therefore, we consider conducting human evaluations. We make comparison with the any-to-any MLLM, NExT-GPT (Wu et al., 2023) that is compatible to multimodal empathetic generation. We prepare 20 dialogue queries from diverse scenarios for two systems to respond. Seven questions from different aspects are used to ask users to evaluate on a Likert scale of 1-100. Figure 5 shows the results, where EmpathyEar is superior to NExT-GPT in all aspects, especially for the emotion consistency of speech and vision.

7 Conclusion

We introduce **EmpathyEar**, a novel, open-source, avatar-based multimodal empathetic chatbot. By employing an LLM at its core, enhanced with multimodal encoders and generators, EmpathyEar supports user inputs from any of text, sound, and vision modalities, and more importantly, producing multimodal empathetic responses, offering users, not just textual responses but also digital avatars with talking faces and synchronized voices. EmpathyEar allows for a richer, more empathetic communication experience, surpassing the limitations of current text-only ERG systems, thus offering emotionally resonant interactions across a broader spectrum of scenarios. The system sets a new standard for human-level empathetic dialogue systems, blending intelligence with the ability to understand and express human emotions.

Limitations and Future Work

Despite the progress `EmpathyEar` makes in empathetic response generation through multimodal integration, there are three main limitations that present opportunities for future work.

First, we rely on external tools for the backend speech generator and talking-head avatar generator, linked to the LLM through text-based commands. This cascading method has inherent limitations; errors in the LLM's output may propagate to the multimodal generation, and the lack of end-to-end learning in our system might restrict performance improvements. Future research could look into developing an integrated end-to-end architecture based on our system.

Second, while our design allows the LLM to produce a meta-response guiding the multimodal generators to maintain consistency in content and emotional tone, there may still be occasional inconsistencies. Investigating methods to enhance cross-modal consistency in semantics and emotional expression could be a focus for further study.

Third, although we introduce the concept of multimodal empathetic response generation, we have yet to define a comprehensive benchmark or standard for this task. Future research should focus on establishing clear definitions, datasets, and validation methods for this area.

Ethics Statement

The development and deployment of `EmpathyEar`, an avatar-based multimodal empathetic chatbot, involve significant ethical considerations. Key concerns include the need to protect user data privacy, particularly emotional data, using strict data protection measures to prevent misuse. It's important to note that `EmpathyEar` does not substitute for professional psychological or medical advice. We commit to the principle of beneficence, aiming to improve user well-being and minimize harm while adhering to ethical standards of fairness, non-discrimination, and bias prevention.

References

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the ICCV*, pages 1708–1718.

Emad Barsoum, Cha Zhang, Cristian Canton-Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction, ICMI 2016, Tokyo, Japan, November 12-16, 2016*, pages 279–283.

Changyu Chen, Yanran Li, Chen Wei, Jianwei Cui, Bin Wang, and Rui Yan. 2024. Empathetic response generation with relation-aware commonsense knowledge. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 87–95.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. 2021. LSSED: A large-scale dataset and benchmark for speech emotion recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*, pages 641–645.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391.

Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024a. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *CoRR*.

Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024b. Enhancing video-language representations with structural

spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. 2023. Efficient emotional adaptation for audio-driven talking-head generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22634–22645.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 807–819.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190.

Akriti Jaiswal, A Krishnama Raju, and Suman Deb. 2020. Facial emotion detection using deep learning. In *2020 international conference for emerging technology (INCET)*, pages 1–5. IEEE.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 119–132.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the ICML*, pages 19730–19742.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2024. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. 2022. MAFW: A large-scale, multi-modal, compound affective database for dynamic facial expression recognition in the wild. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 24–32.

OpenAI. 2022a. Gpt-4 technical report.

OpenAI. 2022b. Introducing chatgpt.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Kashfia Sailunaz, Manmeet Dhaliwal, Jon Rokne, and Reda Alhajj. 2018. Emotion detection from text and speech: a survey. *Social Network Analysis and Mining*, 8(1):28.

Mohammad Soleymani, Sadjad Asghari-Esfeden, Yun Fu, and Maja Pantic. 2015. Analysis of eeg signals and facial expressions for continuous emotion detection. *IEEE Transactions on Affective Computing*, 7(1):17–28.

Linzhuang Sun, Nan Xu, Jingxuan Wei, Bihui Yu, Liping Bu, and Yin Luo. 2023. Rational sensibility: Llm enhanced empathetic response generation guided by self-presentation theory. *arXiv preprint arXiv:2312.08702*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca:

An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. 2020. MEAD: A large-scale audio-visual dataset for emotional talking-face generation. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXI*, volume 12366 of *Lecture Notes in Computer Science*, pages 700–717.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Zhou Yang, Zhaochun Ren, Yufeng Wang, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Yunbing Wu, Yisong Su, Sibo Ju, and Xiangwen Liao. 2024a. Exploiting emotion-semantic correlations for empathetic response generation. *arXiv preprint arXiv:2402.17437*.

Zhou Yang, Zhaochun Ren, Wang Yufeng, Shizhong Peng, Haizhou Sun, Xiaofei Zhu, and Xiangwen Liao. 2024b. Enhancing empathetic response generation by augmenting llms with small-scale empathetic models. *arXiv preprint arXiv:2402.11801*.

Jinfeng Zhou, Chujie Zheng, Bo Wang, Zheng Zhang, and Minlie Huang. 2023. Case: Aligning coarse-to-fine cognition and affection for empathetic response generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8223–8237.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 2022. Emotional voice conversion: Theory, databases and ESD. *Speech Commun.*, 137:1–18.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Specification of Emotion-aware Instruction-Tuning

A.1 Encoder-LLM Alignment Learning

General Alignment Learning: We feed the audio and video into LLM, and then let it output corresponding captions. Datasets include: 'audio-text' AudioCap data (Kim et al., 2019), and 'video-text' Webvid data (Bain et al., 2021).

Emotion-aware Multimodal Alignment: Likewise, we feed the speech audios or videos, and let LLM output the correct emotion labels/types.   Speech-based emotion detection datasets: LSSED (Fan et al., 2021), MELD (Poria et al., 2019); and Vision-based emotion detection data FERPlus (Barsoum et al., 2016) and MAFW (Liu et al., 2022).

Textual Empathetic Response Alignment: Inputting pure textual dialogue contexts encourages LLM to generate correct empathetic response texts. We use the commonly employed text-based EmpatheticDialogue ERG data (Rashkin et al., 2018).

A.2 Meta-response Instruction-Tuning

Following the construction of many existing instruction-tuning datasets (Ouyang et al., 2022; Dai et al., 2023), we also utilize OpenAI GPT-4[5] to generate rich data under the meta-response format defined above. We prompt GPT-4 to fully adhere to the CoT reasoning format, allowing the LLM to simulate this process, where we ensure the data diversification by generating samples evenly covering the pre-setting characters of the avatar, as shown in Table 1.

A.3 Emotion-aware Enhancement Training

While we can directly employ the off-the-shelf well-trained speech generator and talking-head generator for our use, the quality of these two generators might be sub-optimal, especially in terms of their emotional awareness. Thus, we enhance their perceiving by further training them in emotion-aware datasets.   Specifically, we fine-tune the speech generator and talking-head generator on the emotional speech and video dataset, ESD (Zhou et al., 2022) and MEAD (Wang et al., 2020), respectively. We retrofit the ESD and MEAD datasets slightly to meet our requirements. For example, we prepare the speech text by first recognizing the text from the video speech.

---

[5] `https://chat.openai.com/`

Figure 6: Screenshot of the dialogue between a user and EmpathyEar in the child companionship scenario.

Figure 7: Screenshot of the dialogue between a user and EmpathyEar in the health care scenario.
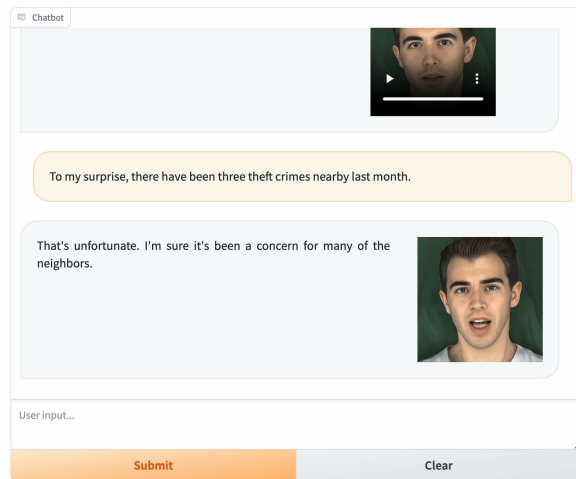

Figure 8: Screenshot of the webpage interface.

B More Demonstrations

Figure 8 shows the system's interactive interface. Figure 6 displays the process of multimodal empathetic responses in the *child companionship* scenario. Figure 7 presents an interactive scenario in the *health care* context.