The ACM Multimedia 2025 Grand Challenge of Avatar-based Multimodal Empathetic Conversation

Han Zhang
Xidian University
Xi'an, China
22021110280@stu.xidian.edu.cn

Hao Fei
National University of Singapore
Singapore, Singapore
haofei37@nus.edu.sg

Hong Han
Xidian University
Xi'an, China
hanh@mail.xidian.edu.cn

Lizi Liao
Singapore Management University
Singapore, Singapore
lzliao@smu.edu.sg

Erik Cambria
Nanyang Technological University
Singapore, Singapore
cambria@ntu.edu.sg

Min Zhang
Harbin Institute of Technology(Shenzhen)
Shenzhen, China
zhangmin2021@hit.edu.cn

Abstract

The Ava-MERG Challenge at ACM Multimedia 2025 explores avatar-based multimodal empathetic response generation by advancing dialogue systems that respond with emotional awareness across text, speech, and talking-face video. To support this goal, we constructed a high-quality dataset featuring aligned multimodal responses in diverse conversational scenarios. The challenge comprises two progressively complex subtasks centering around multimodal-aware empathetic response generation. We aim to promote research in empathetic AI and affective computing. More information is available at https://AvaMERG.github.io/MM25-challenge.

CCS Concepts

• **Human-centered computing → Human computer interaction (HCI)**.

Keywords

Empathetic Response Generation, Affective Computing

ACM Reference Format:
Han Zhang, Hao Fei, Hong Han, Lizi Liao, Erik Cambria, and Min Zhang. 2025. The ACM Multimedia 2025 Grand Challenge of Avatar-based Multimodal Empathetic Conversation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3746027.3762028

1 Introduction

While Large Language Models (LLMs) [2–4, 6, 11, 14, 19, 20] advance machine intelligence, true AGI requires emotional and empathetic abilities. To overcome the limits of text-only Empathetic Response Generation (ERG) [13], we propose the Avatar-based Multimodal Empathetic Response Generation (Ava-MERG) benchmark [5, 21], which requires synchronized responses across text, speech, and facial video. Its accompanying challenge fosters stronger multimodal emotion modeling with applications in assistants, healthcare, and education.

Table 1: Statistics of AvaMERG dataset.

Item		Stats
Dialogue	#Train Set	24,696
	#Valid Set	4,373
	#Test Set	3,979
	#Total	33,048
	Avg. Words Per Utterance	14.68
	Avg. Utterance Per Dialigue	4.6
Modality	Utterance Text	152,021
	Speech Audio	152,021
	Talking-head Video	152,021
	Avg. Length (Sec) Per Aud/Vid	5.67
Avatar	Child (Male/Female)	3/3
	Young (Male/Female)	25/17
	Middle-aged (Male/Female)	4/4
	Elderly (Male/Female)	5/4
	Tone (Emphatic/Mild/Gentle)	14/38/13
	Race	5
Emotion	Text/Multimodal	32/7
Topic&Scenario		10

2 Dataset Specification

We build the AvaMERG dataset by enriching the Empathetic Dialogue (ED) corpus [13] with multimodal and identity annotations. Speaker profiles (age, gender, tone, topic) are generated via GPT-4, and additional dialogues are created to balance demographics. Volunteers then record speech and talking-head videos aligned with these dialogues and profiles.

3 Challenge Overview and Evaluation

This challenge aims to build systems that understand multimodal dialogue contexts and generate emotionally appropriate responses across multimodalities [8–10, 15–18], e.g., text, speech, and video. It consists of two subtasks:

Subtask I: Multimodal-Aware Empathetic Response Generation. Participants develop models that generate empathetic text and speech responses from multimodal dialogue contexts. Evaluation combines automatic metrics—emotion prediction accuracy (**Acc.**) and diversity (**Dist-1/2**)—with human assessments of empathy and fluency (**HE**).

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3762028

Subtask II: Multimodal Empathetic Response Generation. Here, models must additionally produce talking-face videos aligned with generated text and speech. Besides metrics from Subtask I, human judges also evaluate Multimodal Coherence (**MC**), Naturalness (**NT**), and Emotional Expressiveness (**EE**).

Table 2: Task1 Top-performing Results

Rank	Team	Acc. (%)	Dist-1	Dist-2	HE
1	It's MyGO!!!!!	68.42	3.93	23.64	3.8
2	DZ	52.75	3.06	20.17	4.2
3	AI4AI	41.79	4.79	31.79	4.1

Table 3: TASK2 Top-performing Results

Rank	Team	MC	NT	EE	Total
1	SYSU_RUNNER	4.0	3.8	4.3	12.1
2	AI4AI	3.8	4.1	3.6	11.5
3	It's MyGO!!!!!	3.5	3.2	3.5	10.2

4 Top-performing Methods

Tables 2 and 3 present the top-performing results for Subtask 1 and Subtask 2, respectively. Below, we briefly introduce the approaches adopted by these leading methods.

Team It's my Go!!! employs Vicuna [1] to generate empathetic textual responses guided by structured cues and a chain-of-empathy reasoning process. The generated text is then converted into expressive speech using StyleTTS2 [12] and rendered into synchronized facial video through an Emotionally Aligned Talking-face (EAT) [7] generator.

Team SYSU_RUNNER proposes *E3RG*, an Explicit Emotion-driven Empathetic Response Generation system built upon multimodal LLMs. E3RG addresses the AvaMERG task by decomposing it into three stages: empathy understanding, memory retrieval, and response generation, further enhanced by expressive speech and video synthesis modules.

TEAM AI4AI presents *EMO-Avatar*, an emotion-support agent that integrates affective reasoning with multimodal expressiveness. It features a LLM-driven three-stage emotional support strategy and generates empathetic responses across multiple modalities, including voice, gestures, facial expressions, and actions.

TEAM DZ proposes the *MERIA* framework, which leverages a β-VAE-based multimodal disentanglement encoder to effectively mitigate emotional inconsistencies across different modalities and extend the chain of empathy in the AvaMERG dataset to multimodal aspects.

5 Conclusion

We successfully organized the Avatar-based Multimodal Empathetic Conversation Grand Challenge at ACM MM 2025. The challenge attracted broad participation, and the top three teams for each subtask were selected based on their performance. We hope that the AvaMERG Challenge, along with the released benchmark and tools, will foster continued research and innovation in multimodal empathetic response generation.

References

[1] 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%* ChatGPT Quality. https://lmsys.org/blog/2023-03-30-vicuna/

[2] Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*.

[3] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing. *Proceedings of the Advances in neural information processing systems*.

[4] Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. 2024. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

[5] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. Empathyear: An open-source avatar multimodal empathetic chatbot. *arXiv preprint arXiv:2406.15177* (2024).

[6] Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, et al. 2025. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the International Conference on Machine Learning*.

[7] 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*. 22634–22645.

[8] Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.

[9] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*. 18462–18470.

[10] Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. 2023. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5923–5934.

[11] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. 2024. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632* (2024).

[12] 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 (2024).

[13] Hannah Rashkin. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207* (2018).

[14] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605* (2025).

[15] Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14734–14751.

[16] Shengqiong Wu, Hao Fei, and Tat-Seng Chua. 2025. Universal scene graph generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 14158–14168.

[17] Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. 2023. Cross2StrA: Unpaired Cross-lingual Image Captioning with Cross-lingual Cross-modal Structure-pivoted Alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2593–2608.

[18] Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. 2022. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 11513–11521.

[19] 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* (2024).

[20] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the International Conference on Machine Learning*. 53366–53397.

[21] Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards Multimodal Empathetic Response Generation: A Rich Text-Speech-Vision Avatar-based Benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.