arXiv:2601.04960v1 [cs.CL] 8 Jan 2026

面向类人交互、融入注入式情感归因思维的统一口语语言模型

Qing Wang†、*Zehan Li*†、*Yaodong Song*†、*Hongjie Chen*†、*Jian Kang*、*Jie Lian*、*Jie Li*、*Yongxiang Li*、*Xuelong Li**

中国电信人工智能研究院（TeleAI），中国

摘要

本文提出了一种面向情绪智能的统一口语语言模型，并通过一种名为注入式情绪归因思维（IEAT）的新型数据构建策略对其进行增强。IEAT 将用户情绪状态及其潜在原因纳入模型的内部推理过程，使模型能够将情绪感知推理内化，而非将其视为显式监督。模型采用两阶段渐进式策略进行训练。第一阶段通过自蒸馏进行语音-文本对齐和情感属性建模；第二阶段则进行端到端跨模态联合优化，以确保文本与口语情绪表达的一致性。在类人语音对话系统挑战赛（HumDial）情绪智能基准上的实验表明，在基于大语言模型的评估和人工评估中，该方法在情绪轨迹建模、情感推理和共情式响应生成方面均取得了排名领先的表现。

索引词— HumDial，情绪智能，语音对话，语音语言模型

1. 引言

语音对话系统正从面向任务的交互拓展至情绪智能，这要求模型能够理解动态人类情绪并作出回应。然而，在统一的语音-语言框架内实现语义与情绪一致性，仍是一项重大挑战。为此，人类化语音对话系统（HumDial）挑战赛[1] – 赛道一：情绪智能，致力于推动多模态系统中情感推理与回复生成的边界。

在本研究中，我们的系统基于 GOAT-SLM [1] 架构，在 AI Flow 框架 [2]内实现，并融入 BoSS [3]提供的情感线索。系统遵循“聆听–思考–书写–说话”的范式，在共享语义空间中处理语音和文本，并支持连贯生成文本和语音输出。为增强情绪智能，我们提出了注入式情绪归因思维（IEAT），一种数据

一种构建与训练策略，明确将情绪状态及其诱发因素融入模型的内部思维过程。我们的方法在 HumDial 情感智能赛道中排名第一，并在挑战赛全面的评估流程下，于全部三个子任务中展现出卓越且稳定的表现。这些结果凸显了统一

† 共同贡献。

*通讯作者。邮箱：xuelong li@ieee.org

[1] https://github.com/ASLP-lab/Hum-Dial/tree/main

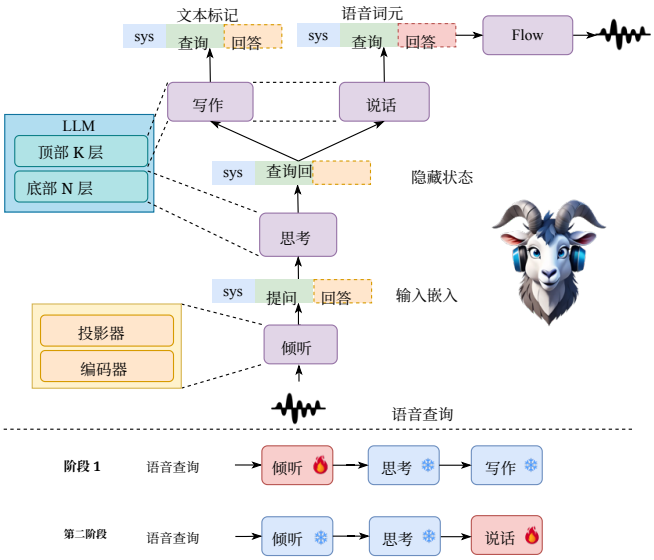

图1。训练流程概述。

融入情绪推理注入的语音-语言建模，是构建具有情感智能的口语对话系统的一条有前景的方向。

2. 系统描述

2.1. 系统概述

我们的系统采用端到端架构，整合了语音感知、语义推理和多模态生成。如图1所示，该模型基于 GOAT-SLM 结构构建，其中语音输入被映射到共享语义空间，并由一个通用推理核心进行处理，从而生成一致的文本和语音输出。

训练采用两阶段渐进式策略。第一阶段通过自蒸馏，着重进行语音-文本对齐和情感归因建模，并逐步从语义对齐扩展到对语言内容、副语言信息和说话人特征的联合建模。第二阶段进行端到端跨模态联合优化，同时训练文本和语音生成路径，以提升不同模态间的情感连贯性和语义一致性。语音生成分支采用模块化微调，对大语言模型顶部的 k 层进行语音标记预测，同时冻结

模型	任务1-中文	任务1-英文	任务2-中文	任务2-英文	任务3-中文	任务3-英文	自动问答-中文 [4]	自动问答-英文 [4]
Freeze-omni [5]	3.11	2.58	3.63	2.79	4.02	3.66	21.23	45.29
Stepaudio2-mini [6]	3.71	2.22	3.81	3.16	4.53	4.30	28.11	47.67
Qwen2.5-omni [7]	4.38	4.40	3.25	3.73	4.28	4.18	22.08	50.13
Qwen3-omni [8]	4.05	3.90	3.80	3.56	4.34	4.31	**49.95**	56.91
Mimo-audio-instruct [9]	4.89	3.92	4.89	4.51	4.00	3.63	34.29	51.63
我们的模型	**4.98**	**4.87**	**4.98**	**4.83**	**4.53**	**4.36**	37.38	**57.69**

表 1。HumDial 挑战赛赛道 1 开发集上的结果。任务*得分遵循官方基于大语言模型的评估标准（0–5 分制 ↑）。音频问答*得分遵循 TELEVAL 标准（0–100 分制 ↑）。最高分以粗体显示，第二高分带下划线。

底部 N 层，以保留基础语言知识。此外，引入多词元预测，以支持实时流式 TTS 合成。

为进一步增强情感理解能力，我们将注入式情感归因思维（IEAT）引入为一种数据构建策略。IEAT 将情绪状态及其潜在原因注入模型的内部推理过程中，使情感信息内化为模型认知的一部分。统一架构、两阶段训练与 IEAT 相结合，使单一口语语言模型能够有效进行情感轨迹建模、情感推理和共情式响应生成。

2.2. 数据构建

我们使用 Qwen3-8B 模型生成训练目标及相应指令。训练集由来自开源数据集 Emilia [10], 、WenetSpeech4TTS [11], 和 GigaSpeech [12] 的中英子集构成。此外，我们从 train 3.5M CN [13, 14] 和 Ultrachat [15], 中抽取合适的用户查询，并使用 CosyVoice2 [16] 模型合成语音查询。随后，我们将这些样本合并到训练集中。在阶段 1-1 中，我们将文本查询或自动语音识别转录文本输入大语言模型，以生成自然的回复，并将其作为音频与文本模态之间跨模态对齐的监督信号。在阶段 1-2 中，为了鼓励模型从声学信息中提取情感线索，而非仅依赖文本语义，我们使用 emo2vec [17] 模型对所有阶段 1-1 的音频样本进行情感分类，为每个样本获取情感标签。随后，我们模仿 Qwen3-8B 在其内部思维模式中采用的分析风格，将情感信息插入思维体，并省略思维终止标记（<\think>），使 Qwen3-8B 继续生成推理轨迹和最终回复。这一过程即为 IEAT 步骤。最终，我们通过应用与三项任务对应的提示，获得阶段 1-2 的输出。为增强模型的多轮对话能力及其在情感轨迹检测任务上的表现，我们抽取所构建共情式响应训练数据的 40%，将其重新整理为 2 轮、3 轮和 4 轮对话，并确保初始与最终情绪状态有所不同。

此外，我们为第一至第二阶段创建了指令音频。通过大语言模型生成后再进行人工筛选，我们为每项任务生成 500 条文本指令，并使用 CosyVoice2 和从 WenetSpeech4TTS 中随机选出的 50 位说话人进行合成。这些指令音频随后与第一至第二阶段的训练音频拼接。共情回应任务不需要额外的指令音频。在第二阶段，我们使用 Emilia 和 LibriHeavy [18] 进行训练，以进一步增强跨模态对齐和语音合成质量。

3. 实验结果

在 TELEVAL 框架 [4], 下，我们在 TELEVAL 中 HumDial 和音频问答（AQA）数据集的官方开发集上评估小语言模型。这使我们能够与多个强大的开源基线系统进行严格比较，包括 freeze-omni [5], 、stepaudio2-mini [6], 、qwen2.5-omni [7], 、Qwen3- omni [8], 和 Mimo-audio-instruct [9]。小语言模型使用默认系统提示词进行推理，所有分数均遵循该挑战赛的官方基于大语言模型的评分指南。如表 1 所示，我们的小语言模型在多语言情感轨迹检测、情感推理和共情回应方面均达到了最先进的情感智能。各项任务和语言上的稳定性能提升，验证了我们统一语音-语言建模方法的有效性，同时保持了常识智能。

Team	任务1	任务2	任务3	Hum	最终得分
我们的方法（第1名）	4.97	4.98	3.85	3.79	4.27
排名第 2	4.90	5.00	4.14	3.70	4.24
排名第 3	4.76	4.76	4.02	3.81	4.21
排名第 4	3.67	4.92	4.93	3.71	4.06
基线	2.62	2.73	2.73	2.96	2.82

表 2。HumDial 挑战赛赛道 1 测试集上的结果。分数遵循官方的基于大语言模型的评估（0–5 分制 ↑）。

测试集上的最终结果如表 2 所示，我们的模型取得了 4.27 的最高最终得分，在所有参赛系统中排名第一。尽管部分竞争方法在个别子任务上取得了相当或略高的分数，但我们的方法在任务 1、任务 2 以及人工评估维度上始终保持强劲性能，从而获得了最佳的总体加权得分，凸显了所提出的结合 IEAT 策略的统一口语语言建模方法的有效性。

4. 结论

在本研究中，我们提出了一种用于情绪智能的统一口语语言模型，并通过注入式情绪归因思维（IEAT）加以增强。通过将情绪状态及其成因融入模型的推理过程，所提出的方法能够在单一框架内有效支持情绪智能。在 HumDial 情绪智能基准上的评测中，我们的方法位居榜首，展现了通过注入情绪推理实现统一口语语言建模的有效性。

5. 参考文献

[1] Hongjie Chen、Zehan Li、Yaodong Song、Wenming Deng、Yi- tong Yao、Yuxin Zhang、Hang Lv、Xuechao Zhu、Jian Kang、Jie Lian 等人，“Goat-slm：具备副语言和说话人特征意识的口语语言模型”，*arXiv* 预印本 *arXiv:2507.18119*，2025年。[2] Hongjun An、Wenhan Hu、Sida Huang、Siqi Huang、Ruanjun Li、Yuanzhi Liang、Jiawei Shao、Yiliang Song、Zihan Wang、Cheng Yuan、Cheng Yuan、Chi Zhang、Hongyuan Zhang、Wen- chao Zhuang 和 Xuelong Li，“AI Flow：观点、场景与方法（2025）”，*arXiv* 预印本 *arXiv:2506.12479*。[3] Qing Wang、Zehan Li、Hang Lv、Hongjie Chen、Yaodong Song、Jian Kang、Jie Lian、Jie Li、Yongxiang Li、Zhongjiang He 等人，“BoSS：超语义语音”，*arXiv* 预印本 *arXiv:2507.17563*，2025年。[4] Zehan Li、Hongjie Chen、Qing Wang、Yuxin Zhang、Jing Zhou、Xuening Wang、Hang Lv、Mengjie Du、Yaodong Song、Jie Lian、Jian Kang、Jie Li、Yongxiang Li 和 Xuelong Li，“Televal：面向中文交互场景中口语语言模型的动态基准”，*arXiv* 预印本 *arXiv:2507.18061*，2025年。[5] Xiong Wang、Yangze Li、Chaoyou Fu、Yunhang Shen、Lei Xie、Ke Li、Xing Sun 和 Long Ma，“Freeze-omni：一种基于冻结大语言模型的智能低延迟语音到语音对话模型”，*arXiv* 预印本 *arXiv:2411.00774*，2024年。[6] Boyong Wu、Chao Yan、Chen Hu、Cheng Yi、Chengli Feng、Fei Tian、Feiyu Shen、Gang Yu、Haoyang Zhang、Jingbei Li 等人，“Step-audio 2 技术报告”，*arXiv* 预印本 *arXiv:2507.16632*，2025年。[7] Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等人，“Qwen2. 5-omni 技术报告”，*arXiv* 预印本 *arXiv:2503.20215*，2025年。[8] Jin Xu、Zhifang Guo、Hangrui Hu、Yunfei Chu、Xiong Wang、Jinzheng He、Yuxuan Wang、Xian Shi、Ting He、Xinfa Zhu 等人，“Qwen3-omni 技术报告”，*arXiv* 预印本 *arXiv:2509.17765*，2025年。[9] LLM-Core-Team Xiaomi，“Mimo-audio：音频语言模型是少样本学习器”，2025年。

[10] Haorui He、Zengqiang Shang、Chaoren Wang、Xuyuan Li、Yicheng Gu、Hua Hua、Liwei Liu、Chen Yang、Jiaqi Li、Peiyang Shi 等，《Emilia：一个用于大规模语音生成的广泛、多语言且多样化的语音数据集》，载于*2024*年*IEEE*口语技术研讨会（*SLT*）。IEEE，2024年，第 885–890 页。

[11] Linhan Ma、Dake Guo、Kun Song、Yuepeng Jiang、ShuaiWang、Liumeng Xue、Weiming Xu、Huan Zhao、Binbin Zhang 和 Lei Xie，《WenetSpeech4TTS：一个包含 12,800 小时普通话语音的语料库，用于大型语音生成模型基准测试》，*arXiv* 预印本 *arXiv:2406.05763*，2024年。

[12] Guoguo Chen、Shuzhou Chai、Guanbo Wang、Jiayu Du、Wei-Qiang Zhang、Chao Weng、Dan Su、Daniel Povey、Jan Trmal、Junbo Zhang 等，《GigaSpeech：一个不断演进的多领域自动语音识别语料库，包含 10,000 小时转录音频》，*arXiv* 预印本 *arXiv:2106.06909*，2021年。

[13] Yunjie Ji、Yan Gong、Yong Deng、Yiping Peng、Qiang Niu、Baochang Ma 和 Xiangang Li，《迈向更优的指令遵循型中文语言模型：探究训练数据与评估的影响》，2023年。[14]BELLEGroup，《Belle：人人都能使用的大语言模型引擎》，https://github.com/LianjiaTech/ BELLE，2023年。[15] Ning Ding、Yulin Chen、Bokai Xu、Yujia Qin、Shengding Hu、Zhiyuan Liu、Maosong Sun 和 Bowen Zhou，《通过扩展高质量指令对话来增强聊天语言模型》，载于*2023*年自然语言处理经验方法会议论文集，2023年，第3029–3051页。[16] Zhihao Du、Yuxuan Wang、Qian Chen、Xian Shi、Xiang Lv、Tianyu Zhao、Zhifu Gao、Yexin Yang、Changfeng Gao、Hui Wang 等，《CosyVoice 2：利用大语言模型实现可扩展的流式语音合成》，*arXiv*预印本 *arXiv:2412.10117*，2024年。[17] Peng Xu、Andrea Madotto、Chien-Sheng Wu、Ji Ho Park 和 Pascale Fung，《Emo2vec：通过多任务训练学习通用情感表征》，*arXiv*预印本 *arXiv:1809.04505*，2018年。[18] Wei Kang、Xiaoyu Yang、Zengwei Yao、Fangjun Kuang、Yifan Yang、Liyong Guo、Long Lin 和 Daniel Povey，《LibriHeavy：包含标点、大小写及上下文信息的5万小时自动语音识别语料库》，2023年。