arXiv:2603.08936v1 [cs.SD] 9 Mar 2026

VoxEmo：基于语音大语言模型对语音情感识别进行基准测试

Hezhao Zhang [1] *Huang-Cheng Chou* [2] *Shrikanth Narayanan* [2] *Thomas Hain* [1]

[1] 英国谢菲尔德大学计算机科学系[2] 南加州大学 Ming Hsieh 电气与计算机工程系信号分析与解释实验室（SAIL），美国加利福尼亚州洛杉矶市，邮编 90089{hzhang181, t.hain}@sheffield.ac.uk

摘要

语音大语言模型（LLM）通过生成式接口在语音情感识别（SER）方面展现出巨大潜力。然而，从封闭集分类转向开放式文本生成会带来零样本随机性，使评估对提示词高度敏感。此外，传统的语音大语言模型基准测试忽视了人类情绪本身的模糊性。因此，我们推出 VoxEmo，这是一个全面的语音情感识别基准，涵盖 15 种语言的 35 个情感语料库，专为语音大语言模型而设计。VoxEmo 提供标准化工具包，涵盖不同复杂度的提示词，从直接分类到副语言推理。为反映真实世界的感知与应用，我们引入了感知分布的软标签协议和提示词集成策略，以模拟标注者之间的意见分歧。实验表明，尽管零样本语音大语言模型在硬标签准确率方面落后于监督基线，但其与人类主观分布的契合度独具优势。

索引词：语音识别、人机交互、计算副语言学

1. 引言

语音情感识别（SER）旨在自动识别语音中表达的情绪与情感 [1, 2]。作为人机交互（HCI）中情感建模的一项关键能力，语音情感识别支持情感感知语音接口和语音分析，例如口语对话系统 [3] 和呼叫中心监控 [4]。尽管近年来语音领域的多项任务取得了显著进展（例如 WavLM [5]），但在自然语音中实现稳健的语音情感识别性能，仍远比在受控表演场景中更具挑战性 [6]。近期，一种应对诸多挑战的新策略应运而生：利用语音大语言模型的多种能力进行语音情感识别 [7]（例如 Qwen-Audio [8]）。语音大语言模型凭借生成式、遵循指令的接口 [9], ，联合利用语音中的词汇语义与副语言线索 [10]，取得了具有竞争力的结果，因此受到广泛关注。

然而，从传统模型向语音大语言模型的转变，与现有的评估范式并不相称。传统上，语音情感识别被定义为一种有监督的封闭集分类任务 [1, 11]。通常，语音编码器用于提取嵌入，轻量级预测头输出类别概率，而推理时行为则由模型架构严格固定 [12, 6]。因此，现有的基准测试框架主要侧重于规范数据划分和评估指标 [13, 6]。相比之下，基于语音大语言模型的语音情感识别作为一种生成式文本输出任务运行。给定一个音频片段和明确的指令/提示词，模型会生成一段文本响应，随后必须对其进行解析并映射到封闭标签集 [14]。因此，报告的得分对推理时协议-高度敏感，具体而言，受到提示词构造、解码设置和解析启发式方法的影响 [15]。这种零样本随机性大幅降低了不同研究之间的可比性，因此亟需制定评估程序，使推理时设置能够得到明确且可复现的记录。

除了语音建模方式的变化之外，SER 基准测试还受到基础数据集语义和领域偏移的混淆影响。情绪本质上具有歧义性和主观性 [16, 17], ，因此标注者间一致性较低，尤其是在自然语料库中 [18, 19]。这意味着需要根据标签来源进行关键区分：标注反映的是说话者意图表达的情绪，还是听者感知到的情绪 [20]。先前研究表明，保留标签分布（软标签）以反映标注和标签不确定性，比将少数意见压缩为单一硬标签更为理想 [21, 22]。此外，随着零样本语音大语言模型实现无缝的跨语料库评估，不加引导的比较往往会混淆多个变量，例如语言、互动风格或录音条件 [23, 24]。因此，适当的基准应当通过明确且可解释的偏移类型，对跨语料库迁移性能进行分类。

总而言之，生成模型随机性和数据集语义模糊性的影响要求建立一个统一的大规模基准测试框架。我们推出 VoxEmo，这是一款评估工具包和基准，旨在为基于语音大语言模型的语音情感识别研究提供标准。我们的贡献包括以下四点：

- (i) 一款标准化的基于语音大语言模型的语音情感识别评估工具包，统一了提示词模板、生成设置、输出解析和无效输出处理。
- (ii) 一份全面的基准报告和排行榜，涵盖 15 种语言的 35 个语料库，并附有一份披露清单，以确保比较结果可复现。
- (iii) 一种新颖的数据集元数据模式，明确记录**标签来源**（感知的与表达的），以支持分布感知评估。
- (iv) 结构化、由语料库驱动的跨领域设置，明确区分标签集偏移和声学偏移。

2. VoxEmo 基准设计

为满足上述要求，VoxEmo 整理了 35 个带有明确元数据并采用标准化划分的语料库。

2.1. 数据概述

表 1 总结了 VoxEmo 使用的数据集，并展示了各数据集的主要属性。如表所示，纳入了 2006 年至 2025年间发布的 35 个数据集，涵盖 15 种语言。

表 1： 数据集概览，按表演式录音与真实场景录音分组，并按年份排序。音频来源（音频来源）报告真实场景语料库的录音来源，以及表演式语料库的脚本化、自发或混合（脚本/即兴）风格。标签来源表示可用的标签来源（标签来源）：*Expr*（表达的/演员意图），*Perc*（感知的/听者标注者）。*Emo*、*#Utts* 和 *#Hrs* 反映经过数据集特定筛选和标签集限制后的基准子集（例如，排除 *MSP-Podcast* 中的无一致样本，并采用 *IEMOCAP* 中普遍使用的标签子集）。

数据集	Year	音频来源	标签来源	#情感	#说话人	选定的语句数	#Hrs	语言
真实场景								
EmotionTalk [25]	2025	脱口秀	Perc	7	30	19,250	23.6	普通话
MSP-Podcast [19, 18]	2025	播客	Perc	8	3,711	212,801	324.1	英语
BIIC-Podcast [26]	2023	播客	Perc	8	–	69,874	147.2	台湾普通话
M3ED [27]	2022	TV	Perc	7	634	24,437	9.8	普通话
ASVP-ESD [28]	2020	在线媒体	Perc	14	132	13,964	18.0	英语/普通话
MELD [29]	2019	TV	Perc	7	–	13,703	12.0	英语
URDU [30]	2018	脱口秀	Perc	4	38	400	0.3	Urdu
表演								
EmoDB [31, 32]	2025	脚本化	Expr	7	10	816	0.6	德语
Thai-ser [33]	2025	脚本/即兴	Expr	5	200	27,854	41.6	Thai
nEMO [34]	2024	脚本化	Expr	6	9	4,481	3.1	波兰语
ASED [35]	2023	脚本化	Perc	5	65	2,474	2.1	阿姆哈拉语
EMNS [36]	2023	脚本化	Expr	8	1	1,181	1.9	英语
Emozionalmente [37]	2023	脚本化	Expr	7	431	6,902	6.3	意大利语
RESD [38]	2022	未知	Expr	7	–	1,337	2.1	俄语
ESD [39]	2021	脚本化	Expr	5	20	35,000	29.1	英语/普通话
MESD [40]	2021	脚本化	Expr	6	16	862	0.2	西班牙语
MEAD [41]	2020	脚本化	Expr	8	60	31,729	37.3	英语
Oreau [42]	2020	脚本化	Expr	7	32	434	0.3	法语
波兰语 [43]	2020	脚本化	Expr	3	5	450	0.1	波兰语
TurEV-DB [44]	2020	脚本化	Expr	4	6	1,735	0.5	土耳其语
ShEMO [45]	2019	自发	Perc	6	87	2,838	3.3	波斯语
SUBESCO [46]	2019	脚本化	Expr	7	20	7,000	7.8	孟加拉语
AESDD [47]	2018	脚本化	Expr	5	6	604	0.7	希腊语
CaFE [48]	2018	脚本化	Expr	7	12	936	1.1	法语
EmoV-DB [49]	2018	脚本化	Expr	5	4	6,887	9.5	英语
JL-Corpus [50]	2018	脚本化	Expr	10	4	2,400	1.4	英语
RAVDESS [51]	2018	脚本化	Expr	8	24	1,440	1.5	英语
CREMA-D [52]	2014	脚本化	表情+Perc	6	91	7,442	5.3	英语
EMOVO [53]	2014	脚本化	Expr	7	6	588	0.5	意大利语
SAVEE [54]	2014	脚本化	Expr	7	4	480	0.5	英语
PAVOQUE [55]	2013	脚本化	Expr	5	1	7,334	12.2	德语
TESS [56]	2010	脚本化	Expr	7	2	2,800	1.6	英语
CASIA [57]	2008	脚本化	Expr	6	6	1,200	0.6	普通话
IEMOCAP [20]	2008	脚本/即兴	表情+Perc	4	10	5,531	7.0	英语
eNTERFACE [58]	2006	自发	Expr	6	44	1,263	1.1	英语

数据集分为真实场景录音（7 个）和表演式录音（28 个），并按发布年份排序。表中列出了发布年份、语言、情感类别数量、说话人和语句数量、语音时长，以及 **音频来源** 和 **标签来源**。**音频来源** 表示录音属于“真实场景”，还是表演式数据集中的脚本式/自发式风格。**标签来源** 用于区分标注针对的是感知情感、表达情感，还是两者兼有；研究发现，真实场景数据集主要提供感知标签，而表演式数据集通常提供表达标签，只有少数同时提供表达标签和感知标签。

2.2. 软标签真值

在由多位听者标注的数据集中，标注者之间的分歧往往反映了有意义的感知差异，而非噪声 [59]。为保留这一信息，真值被建模为未经平滑的、基于计数的**软标签分布 y**，其中表示定义在 C 情感类别上的概率单纯形。形式上，给定某个样本的一组 N 标注，目标分布定义如下：

$$\mathbf{y} = \left[\frac{n_1}{N}, \frac{n_2}{N}, \ldots, \frac{n_C}{N}\right], \tag{1}$$

其中， n_c 表示类别 c获得的投票数。例如，在一个 4 类场景 {中性、愤怒、悲伤、高兴}，若有$N = 5$ 名标注者，其分布为 $\{1, 2, 2, 0\}$，则所得的标签向量为—

表 2： 对比评估的两个语音大语言模型。

	Qwen2-Audio	Audio Flamingo 3
音频编码器	Whisper-large-v3	Whisper-large-v3
大语言模型骨干	Qwen-7B	Qwen2.5-7B
训练数据	37 万小时，未公开	>9M 个问答对，已发布
IEMOCAP WA（%）	59.2	63.8

$\mathbf{y} = [0.2, 0.4, 0.4, 0.0]$。由于并非所有公开情绪数据库都提供标注者级评分，因此软标签分析仅限于包含此类元数据的五个数据集：CREMA-D、IEMOCAP、MSP-Podcast、BIIC-播客和 EmotionTalk。

2.3. 域内划分

遵循 EmoBox 协议 [13], ，该协议为大量公开情感识别语料库提供可复现的说话人独立划分。对于具有提供方定义划分的数据集（例如 IEMO-CAP、MELD），直接采用这些划分。对于其余数据集，划分策略由说话人数以及不同说话人之间情绪标签的平衡情况决定。说话人少于 4 位，或不同说话人之间的情绪标签分布极不均衡的数据集，采用分层随机抽样进行划分（75% 用于训练，25% 用于测试；一折），并按情绪类别分层。对于有 4–6 位说话人的数据集，采用留一说话人交叉验证，折数与不同

表 3：*VoxEmo* 中使用的零样本提示模板。提示类别通过向共享模板添加指令块构成。对于硬标签评估，将最终标签解析为预测。

零样本提示模板	
任务描述	你是一个语音情感识别系统。 你收到了一段音频片段。
情感类别	硬标签提示：恰好选择一个 来自此封闭集合的情感标签：{标签}。 分布提示：考虑以下 情绪标签：{标签}。
指令块	
+ 直接决策	直接根据音频判断情绪。
+ 分布	根据以下内容估计每种情绪的可能性 音频。为每个标签分配一个概率（介于 0.0 和 1.0 之间）。概率必须 总和为 1.0。同时说明哪一个单独的标签是 最有可能。
+ 自动语音识别转录文本	尽可能准确地转录音频中的口语内容 尽可能准确地。
+ 声学描述	描述你听到的声学和副语言线索 听到的内容（例如音高、响度、语速、 停顿、声音质量、非语音事件）。
+ 推理	解释转录文本内容和 声学线索如何共同支持你的最终选择。
响应格式	[自动语音识别转录文本：<transcript>] [声学描述： <描述>] [推理：<reasoning>] [情绪分布：- {“标签1”：概率1, “标签2”：概率2, ...}] 最终标签：<来自一个标签 {标签}>

说话人的数量相同。对于说话人超过六位的数据集，将说话人划分为四折。所有基于说话人的方案都确保训练集和测试集之间没有说话人重叠。对于缺少预定义/官方验证集的数据集，通过分层抽样从训练划分中留出 20% 作为验证集。

2.4. 跨域配置

为评估英语数据中的跨语料库泛化，我们定义了一种跨场景设置，其中标签集、录音条件和**标签来源**（表达的与感知的）同时变化。与传统分类器不同，生成式语音情感识别可以在推理时使用与训练或微调时不同的提示指定标签集，从而便于评估这些变化带来的影响。我们评估了四个语言相同、但标签集大小（4–8 类）、音频来源和标签来源存在系统差异的语料库，并涵盖所有源–目标组合：CREMA-D（6 类、表演、表达标签）、IEMOCAP（4 类、表演式对话、感知的）、MELD（7 类、广播、感知的）和 MSP-Podcast v2.0（8 类、播客、感知的）。发布的工具包还包含跨语言配置；受篇幅所限，本文不作报告。

3. 实验设置

3.1. 语音大语言模型的选择

为最大限度地确保可复现性，同时尽量减少实现驱动的方差，该基准聚焦于能够接受音频输入并在指令提示下生成文本响应的开放权重语音大语言模型。我们选取了两个具有代表性的 ∼7B 音频语言模型：Qwen2-Audio-7B-Instruct（Q2A） [8] 和 Audio Flamingo 3（AF3） [60]。表 2 总结了它们的主要属性。

Q2A 的训练数据组成未公开披露，因此难以评估潜在的数据污染。AF3 的已发布的训练数据明确包含 IEMOCAP 和 MELD 等与语音情感识别相关的语料库，证实其与基准数据集存在重叠。表 2 中的 IEMOCAP 加权准确率分数由 [60]报告。两种模型均由 Hugging Face Transformers[1] 和 PEFT[2], 支持，可在统一且一致的环境中进行推理和微调。

3.2. 零样本协议

零样本评估遵循固定的推理时协议：生成采用贪心解码，所有运行均使用相同的输出解析规则；只有提示模板有所变化。为评估辅助推理模式如何影响语音情感识别性能，我们通过将指令块逐步组合到共享模板中，构建了五种提示变体（表 3）：自动语音识别转录文本（**T**）、声学描述（**A**）和推理（**R**）。对于五个多标注者语料库上的软标签评估（第 2.2 节），每种变体还会额外包含 **分布**块，该模块要求给出情感类别上的概率分布。

对于硬标签评估，预测标签从最终标签字段中提取，并通过不区分大小写的匹配映射到数据集特定的封闭标签集。无法映射其输出的样本被保留为错误预测。在软标签评估中，情绪分布字段被解析为JSON对象，并通过别名规范化映射到目标标签集，然后重新归一化，使概率之和为一。当解析过程失败时，假定所有标签上为均匀分布。

由于单提示词评估对提示词措辞和解析失败较为敏感，因此采用基于投票的集成，将5个硬标签预测聚合为单一分布。这遵循了近期研究发现，即多种提示词引出的预测能更好地反映人工标注分布 [61]。对于样本 k，令$n_{k,c}$ 为预测类别 c的提示词数量， f_k 为解析失败次数；每次失败贡献 $1/C$的均匀投票。预测分布为：

$$\hat{P}(k,c) = \frac{n_{k,c} + f_k/C}{N_{\text{prompts}}}, \quad N_{\text{prompts}} = 5. \tag{2}$$

3.3. 监督微调

为提供与零样本评估相对照的监督参考，我们对 3.1 节中介绍的两个模型进行适配（微调），使其适用于训练集。遵循近期基于 SpeechLLM 的语音情感识别实践，我们将 LoRA [62], 适配器 [63] （$r = 8, \alpha = 16$，dropout 0.05）应用于语言解码器的注意力投影矩阵（q/k/v/o）。音频编码器和其余参数保持冻结，这意味着更新的参数少于总参数的 0.15%。训练使用 AdamW，学习率为 1×10^{-5}，预热比例为 10%，采用 bf16精度，有效批量大小为 16，共训练 10 个 epoch。与多区块零样本模板相比，监督微调（SFT）采用单一的封闭集指令提示，其中注入数据集专属的标签集 {标签}；训练目标的格式为 最终标签：<标签>。推理采用贪婪解码，与零样本设置保持一致。预测结果使用与零样本运行相同的评估器和评估指标进行评分。

3.4. 评估指标

对于硬标签情形（每个话语仅对应一种情绪），报告的标准评估指标包括加权准确率（WA）、未加权准确率（UA）、

[1]https://github.com/huggingface/transformers[2] https://github.com/huggingface/peft

Micro-F1 和宏平均 F1，均针对每个数据集的封闭标签集计算。然而，这些评估指标无法反映多标注者语料库中存在的标注者分歧或情绪的自然歧义性（每个话语可能对应多种情绪）。遵循 [21], ，我们还采用双指标方法报告软标签指标：

• **硬决策评估：**对预测概率分布进行阈值化处理，以计算分类性能，并报告宏平均 F1、Micro-F1 和 Top-1准确率。

• **分布感知评估：**使用 Kullback-Leibler散度（KLD）、Jensen-Shannon散度（JSD）和总变差距离（TVD）衡量概率散度；余弦相似度（Sim）和均方误差（MSE）分别用于评估方向对齐和重建误差。

4. 结果与分析

4.1. 零样本评估

表 4–5 展示了 Q2A 和 AF3 在 35 个数据集上采用五种零样本提示变体时的宏平均 F1、UA 和 WA 结果。可以看出，在几乎所有情况下，提示选择都会对性能产生显著影响。对于 Q2A，在 11 个数据集上，表现最佳与最差的变体之间的宏平均 F1 差距超过 20，在 TESS 上差距最大，达到 49。最佳变体因模型和数据集而异：AF3 在 25/35 个数据集上使用直接提示时表现更佳，而 Q2A 最常受益于声学描述变体（+A，18/35）。从绝对值来看，零样本性能仍然有限：两个模型在八个数据集上的最佳提示宏平均 F1 均低于 20%，且在所有 30 个可比较数据集上都不及 EmoBox监督基线。在最佳提示对比下，Q2A 在 35 个数据集中的 26 个上优于 AF3，不过这一优势应结合第 3.1 节讨论的训练数据重叠加以解读。以下各小节将进一步分析各个指令块如何导致这种差异。

4.1.1. AcousticCaption 的效果

对于 Q2A，声学描述提示（+A）最常取得最高的宏平均 F1，在 35 个数据集中的 18 个上表现如此。不过，这一优势主要集中于表演式语料库（28 个表演数据集中的 16 个，相比真实场景数据集 7 个中的 2 个）。对于脚本化录音，不同情感类别的转录文本内容是固定的，因此副语言线索成为主要的区分信号。明确引导模型生成声学描述，似乎有助于模型关注这些线索。在表演子集中，这一收益主要体现在非英语数据上（18 个非英语表演数据集中的 13 个，相比英语表演数据集 10 个中的 3 个），这表明副语言描述不依赖具体语言的特性，还能在一定程度上弥补较弱的跨语言文本理解能力。AF3 的响应明显较弱：直接提示在 35 个数据集中的 25 个上仍是最佳变体，而 +A 在 35 个数据集中的 28 个上相较于直接提示降低了宏平均 F1。两种模型之间的差异，可能源于训练数据构成不同，或源于将声学描述整合到语言解码器中的方式不同。

*4.1.2. ASR*转录文本的影响

加入转录文本提示（+T）会降低两个模型在大多数表演式数据集上的宏平均 F1（Q2A 为 22/28，AF3 为 25/28）；其中，Q2A 的宏平均 F1 在 CREMA-D 上最多下降 46.7，在 TESS 上下降 37.6。正如第 4.1.1 节所述，脚本化录音在不同情绪类别间固定了转录文本内容，因此新增的词汇信号不具备区分性，并可能引入噪声。

在真实场景语料库中，Q2A 的情况恰好相反：其在 7 个数据集中的 6 个上均有所提升，唯一的例外是 URDU。这些增益主要集中在英语和普通话上，这两种语言在 Q2A 的训练数据中所占比例最高；URDU 的结果则与当前语音大语言模型对代表性不足的语言覆盖有限这一点相一致。相比之下，AF3 在任何真实场景数据集上都未从 +T中获益，7 个变化值全部为负，包括 MELD 等英语语料库。尽管 AF3 的训练数据以英语为主，但其在英语语料库上的这一失败表明，问题在于模型层面利用转录文本内容进行情绪区分的能力受限，而不仅仅是语言覆盖问题。表演式与自然场景数据之间 +T有效性的分化，与表 1 中的 **音频来源** 元数据直接吻合，证实提示敏感性系统性地受到语料库构建方式的影响。

4.1.3. 组合提示词

将转录文本与声学描述合并到单个提示词中（+T+A），似乎并未提升任一模型在 35 个数据集上的宏平均 F1。对于 70 组模型—数据集组合，最优方案在 66 种情况下是单块提示或直接提示（Q2A 为 32/35，AF3 为 34/35）。加入一个推理步骤（+T+A+R）部分恢复了 Q2A 的性能，在 28/35 个数据集上优于 +T+A，但仅在 3 个数据集上超过了最佳单块提示的表现。因此，在这一基准中，单块提示始终优于多块替代方案。这种模式或许反映了两种拥有 70 亿参数的模型在整合多个信息源时受到容量限制，不过对于规模更大或训练方式不同的语音大语言模型，这一结论未必成立。

4.2. 监督微调

表4–5中的监督微调行报告了 Q2A 和 AF3 在 LoRA微调（第 3.3 节）后的测试宏平均 F1。监督微调在 36 个评估划分中的 34 个上，使 Q2A 的表现优于其最佳零样本提示，宏 F1 平均绝对提升为 23.7。在 9 个数据集上，提升超过 40 分，其中包括零样本宏平均 F1 低于 25 的情况。与自然场景数据相比，表演式语料库上的提升更为显著，监督微调的宏平均 F1 均值分别达到 66.0 和 40.1。只有 EmoDB（−2.9）和 URDU（−0.5）略有下降。

在 30 个具有可比 EmoBox 监督基准的数据集上， [13], Q2A 有 10 个数据集的表现超过该基准，另有 5 个数据集的宏 F1 与其相差不超过 3 分（例如 TESS、ESD、eNTER-FACE）。这一表现与数据集规模相关：在至少包含 5,000 条语句的 11 个数据集中，Q2A 有 7 个超过该基准；而在少于 1,000 条语句的 9 个数据集中，则没有一个超过该基准。在自然场景语料库中，Q2A 在 MELD（+6.3）和 M3ED（+4.4）上超过该基准；这两个数据集的 EmoBox 监督基线本身仍低于 34 宏平均 F1，表明语音大语言模型微调在自然语音数据上具有竞争力，而传统方法在这些数据上也面临困难。尽管如此，监督微调后仍有 10 个评估划分的宏平均 F1 低于 40。5 个表演式案例均为语句少于 1,500 条的语料库（EMNS、RESD、EMOVO、MESD、Polish），训练数据有限可能限制了 LoRA 适配带来的提升。5 个自然场景案例涵盖的语料库规模范围更广，从 URDU（400 条语句）到 MSP-Podcast（21.3 万条），这表明除数据量之外，还有其他因素影响自然语音任务的难度。

在相同的 LoRA 配置下，AF3 在 36 个划分中的 26 个上有所提升，但平均增益为 10.3，不到 Q2A 的 23.7 的一半。其余 10 个划分的表现有所下降，其中降幅最大

表 4： *VoxEmo* 零样本 ＋监督微调结果。*Q2A* =*Qwen2-Audio*、*AF3* = *Audio Flamingo 3*。提示词缩写：*T* = 自动语音识别转录文本，*A* = 声学描述，*R* = 推理。 红色粗体：每个数据集的最佳 *Q2A* 提示词； 红色粗体：每个数据集的最佳 *AF3* 提示词。 红色粗体：每个数据集中 *Q2A* 监督微调 / *AF3* 监督微调 / *EmoBox* 之间的最佳分数（按宏平均 *F1*）。第 *1/2* 部分。

设置	UA(%)	WA(%)	F1分数(%)	UA(%)	WA(%)	F1分数(%)	UA(%)	WA(%)	F1分数(%)	UA(%)	WA(%)	F1分数(%)
	AESDD (el, 5)			ASED (am, 5)			ASVP-ESD (英语/zh, 14)			BIIC-Podcast (繁体中文, 8)		
Q2A 直接	28.57	28.19	20.06	35.93	35.39	32.74	11.28	16.61	9.91	27.73	29.85	18.34
Q2A +T	28.18	28.19	27.08	24.59	25	18.34	7.84	11.27	10.38	**28.66**	**33.7**	**19.96**
Q2A +A	**39.52**	**39.6**	**38.63**	**41.41**	**42.21**	**39.02**	**20.86**	**24.55**	**18.4**	21.59	39.35	16.55
Q2A +T+A	27.38	27.52	24.67	29.11	29.87	25.16	9.45	12.47	10.17	20.51	36.98	15.29
Q2A +T+A+R	26.92	26.85	25.77	30.5	31.17	26.81	14.94	21.08	15.47	24.3	40.03	18.08
AF3 直接	29.45	29.53	24.95	**37.18**	**38.15**	**33.28**	**28.79**	**34.41**	**28.36**	**18.26**	**41.22**	**15.43**
AF3 +T	**29.47**	**29.53**	**26.65**	24.87	25.65	17.72	15.55	19.44	16.84	16.05	37.87	13.41
AF3 +A	26.05	26.17	20.98	31.69	32.63	25.39	28.18	36.42	28.03	20.2	36.15	15.19
AF3 +T+A	18.05	18.12	11.97	25.71	26.62	18.25	18.65	24.75	19.27	15.07	39.69	12.58
AF3 +T+A+R	24.09	24.16	21.98	23.48	23.86	18.59	17.53	22.97	19.38	13.54	35.81	10.67
Q2A 监督微调	44.3	44.3	40.09	83.99	84.09	83.96	59.03	70.38	59.51	**39.5**	**57.9**	**40.96**
AF3 监督微调	35.43	35.57	29.72	67.77	67.69	67.34	45.76	58.02	47.7	22.79	48.13	24.25
EmoBox[13]	**84.4**	**84.49**	**84.19**	**96.75**	**96.73**	**96.74**	**61.14**	**71.52**	**62.08**	–	–	–
设置	CaFE（法语，7）			CASIA（中文，6）			CREMA-D（英语，6）			EMNS（英语，8）		
Q2A 直接	37.1	38.57	34.12	31.25	31.25	25.31	**65.94**	**65.3**	**66.22**	14.02	13.4	7.21
Q2A +T	28.67	27.24	25.22	20.08	20.08	14.49	25.71	23.87	19.5	**12.46**	**12.72**	**7.42**
Q2A +A	**40.58**	**38.14**	**37.62**	**40.5**	**40.5**	**35.51**	63.5	62.82	62.14	12.75	13.06	6.9
Q2A +T+A	40.08	36	35.41	26.75	26.75	20.71	32.95	31.52	32.61	13.19	13.4	5.01
Q2A +T+A+R	35.32	32.16	32.27	29.08	29.08	22.43	41.79	40.36	39.43	12.15	12.37	3.98
AF3 直接	**37.5**	**34.62**	**31.25**	29.75	29.75	25.31	**58.49**	**57.49**	**59.78**	**22.58**	**21.99**	**18.54**
AF3 +T	29.27	24.14	24.82	21	21	13.45	38.98	37.5	39.62	13.94	14.09	6.31
AF3 +A	32.34	27.78	27.6	**31.83**	**31.83**	**28.49**	50.36	49.35	50.81	16.57	16.5	10.4
AF3 +T+A	29.07	23.72	24.11	24.42	24.42	18.33	37.15	35.67	37.14	11.62	11.68	5.27
AF3 +T+A+R	27.28	22.44	20.91	23.42	23.42	17.42	28.98	27.59	28.78	9.2	9.28	4.08
Q2A 监督微调	65.58	65.06	64.66	**60**	**60**	**56.57**	**91.17**	**91.06**	**91.11**	14.41	14.43	11.08
AF3 监督微调	38.89	38.14	37.17	46.33	46.33	44.44	74.75	74.84	75.2	21.46	21.31	22.64
EmoBox[13]	**69.43**	**68.84**	**68.06**	59.58	59.58	56.27	76.75	76.48	76.6	**83.97**	**84.12**	**83.97**
设置	EmoDB（德语，7）			EmotionTalk（中文，7）			EmoV-DB（英语，5）			EMOVO（意大利语，7）		
Q2A 直接	69.01	67.97	64.57	27.87	17.32	19.55	**68.04**	**64.05**	**62.1**	23.64	23.64	19.66
Q2A +T	39.14	33.77	36.68	35.87	38.21	27.22	40.67	39.03	37.13	16.33	16.33	12.77
Q2A +A	**75.94**	**68.4**	**70.87**	29.08	37.79	22.44	57.53	58.17	55.71	**27.04**	**27.04**	**21.18**
Q2A +T+A	43.89	40.69	42.53	28.17	45.77	26.27	33.79	33.86	30.76	22.28	22.28	16.67
Q2A +T+A+R	47.33	42.86	45.71	**30.41**	**47.07**	**28.02**	42.42	41.24	38.23	23.81	23.81	18.55
AF3 直接	**54.75**	**54.11**	**53.71**	**28.13**	**45.31**	**27.64**	**74.5**	**75.97**	**77.49**	25.17	25.17	16.35
AF3 +T	47.73	45.45	48.61	22.18	46.5	24.53	60.83	62.13	65.7	17.69	17.69	10.13
AF3 +A	49	48.92	51.11	24.49	45.31	22.98	65.14	66.9	69.1	**24.32**	**24.32**	**16.9**
AF3 +T+A	44.4	42.42	46.94	21.38	49.35	23.04	51.02	52.82	56.19	18.2	18.2	10.74
AF3 +T+A+R	37.65	37.66	41.17	18.24	42.72	19.36	45.8	47.18	51.15	18.71	18.71	11.78
Q2A 监督微调	72.59	73.59	68	**59.19**	**71.28**	**60.22**	**99.88**	**99.88**	**99.88**	33.33	33.33	30.46
AF3 监督微调	33.38	32.03	30.7	38.86	58.68	41.3	89.8	90.29	90.45	16.84	16.84	11.23
EmoBox[13]	**92.58**	**92.67**	**92.57**	–	–	–	99.44	99.47	99.45	**57.82**	**57.82**	**56.06**
设置	Emozionalmente（意大利语，7）			eNTERFACE（英语，6）			ESD（英语/中文，5）			IEMOCAP（英语，4）		
Q2A 直接	23	23	17.84	43.27	43.27	40.09	38.18	38.18	35.05	54.38	56.97	51.63
Q2A +T	18.99	18.99	14.46	**63.14**	**63.14**	**60.26**	25.66	25.66	21.52	44.58	46.9	46
Q2A +A	**27.12**	**27.12**	**22.32**	43.91	43.91	38.09	**42.07**	**42.07**	**40.3**	**57.33**	**59.09**	**58.19**
Q2A +T+A	21.6	21.6	16.98	53.53	53.53	52.36	28.47	28.47	24.3	39.92	43.85	40.03
Q2A +T+A+R	23.23	23.23	18.55	56.09	56.09	55.19	33.96	33.96	30.06	41.25	44.9	41.81
AF3 直接	**24.56**	**24.56**	**20.72**	**53.53**	**53.53**	**56.12**	37.65	37.65	32.34	**62.6**	**63.88**	**63.97**
AF3 +T	16.26	16.26	8.51	48.72	48.72	49.05	28.43	28.43	22.66	46.57	48.88	49.9
AF3 +A	22.01	22.01	15.35	49.04	49.04	48.68	**36.69**	**36.69**	**34.15**	60.1	61.4	61.09
AF3 +T+A	16.9	16.9	10.22	44.55	44.55	46.11	28.31	28.31	23.28	39.08	42.07	41.11
AF3 +T+A+R	15.45	15.45	9.14	38.14	38.14	39.4	24.58	24.58	21.26	43.11	45.99	43.38
Q2A 监督微调	74.62	74.62	74.61	95.83	95.83	95.85	83.7	83.7	83.58	**82.65**	**82.13**	**82.43**
AF3 监督微调	54.65	54.65	54.28	60.9	60.9	61.19	71.72	71.72	71.75	63.46	65.01	65.15
EmoBox[13]	**76.91**	**76.91**	**76.9**	**97.69**	**97.68**	**97.68**	**84.62**	**84.62**	**84.33**	73.54	72.86	73.11
设置	JL-Corpus（英语，10）			M3ED（中文，7）			MEAD（英语，8）			MELD（英语，7）		
Q2A 直接	**20.79**	**20.79**	**14.77**	17.18	20.13	11.38	**16.14**	**16.76**	**13.38**	31.68	42.91	25.97
Q2A +T	12.88	12.88	9.52	**21.49**	**36.95**	**18.93**	16.05	14.24	13.22	**32.23**	**42.98**	**28.06**
Q2A +A	16.46	16.46	9.68	18.63	27.51	12.81	15.1	11.04	8.88	29.07	47.39	24.81
Q2A +T+A	11.08	11.08	4.19	15.39	37.99	12.97	14.71	9.88	8.65	27.37	46.24	25.49
Q2A +T+A+R	12.96	12.96	8.28	17.93	40.71	15.2	14.06	9.41	8.28	27.75	48.04	26.48
AF3 直接	**46.5**	**46.5**	**49.06**	18.25	33.85	14.88	**15.81**	**11.17**	**8.92**	**29.77**	**44.75**	**29.83**
AF3 +T	22.17	22.17	21.82	14.54	40.28	12.38	13.79	8.06	5.48	25.53	49.16	27.52
AF3 +A	32.58	32.58	33.25	**19.23**	**39.47**	**15.04**	14.87	9.37	7.56	27.82	47.85	27.91
AF3 +T+A	21.67	21.67	21.48	16.05	42.83	13.04	12.66	6.85	3.83	20.61	49.46	21.63
AF3 +T+A+R	20.88	20.88	20.28	15.23	40.26	12.91	12.58	6.96	4.3	22.74	50.15	23.39
Q2A 监督微调	44.33	44.33	41.83	**37.57**	**54.95**	**38.24**	69.36	70.26	69.63	**37.85**	**56.33**	**39.26**
AF3 监督微调	37.08	37.08	37.09	25.91	50.24	26.78	48.39	49.36	48.94	24.12	48.54	25.74
EmoBox[13]	**66.71**	**66.71**	**65.19**	32.84	49.42	33.76	**81.27**	**82.03**	**81.43**	31.54	51.89	32.95
设置	MESD（西班牙语，6）			MSP-Podcast(test1)（英语，8）			MSP-Podcast(test2)（英语，8）			nEMO（波兰语，6）		
Q2A 直接	23.27	23.36	17.75	22.23	17.2	12.28	17.99	16.53	9.18	21.88	22.43	14.3
Q2A +T	22.62	22.43	19.74	**24.84**	**31.42**	**20.35**	**19.6**	**32.18**	**15.35**	15.38	16.18	11.35
Q2A +A	**21.45**	**21.5**	**20.54**	21.79	38.1	16.89	16.86	55.48	14.69	**21.34**	**22.07**	**16.89**
Q2A +T+A	20.74	20.56	17.65	20.91	35.59	17.72	15.99	43.15	13.32	17.22	18.14	12.36
Q2A +T+A+R	18.85	18.69	15.56	20.62	37.19	18.78	16.05	45.54	14.13	17.66	18.59	13.17
AF3 直接	**26.5**	**26.17**	**22.32**	**29.11**	**47.66**	**28.67**	**20.58**	**51.42**	**18.78**	**17.88**	**19.12**	**9.04**
AF3 +T	16.59	16.36	12.69	21.35	38.34	18.77	15.7	51.26	13.3	15.45	16.71	5.5
AF3 +A	21.81	21.5	15.1	23.57	42.34	21.38	18.4	56.36	16.57	17.57	18.86	7.75
AF3 +T+A	17.12	16.82	6.84	16.87	35.99	14.81	14.4	52.78	12.2	16.03	17.34	5.71
AF3 +T+A+R	17.59	17.29	7.65	16.29	34.06	14.33	13.05	47.07	10.72	15.54	16.8	5.67
Q2A 监督微调	34.29	34.11	34.3	**32.08**	**45.3**	**29.18**	**23.24**	**48.26**	**18.21**	**78.16**	**78.28**	**78.07**
AF3 监督微调	36.2	35.98	31.62	22.01	38.54	19.08	18.75	48.89	14.86	61.86	62.2	61.56
EmoBox[13]	**69.78**	**69.67**	**69.64**	–	–	–	–	–	–	–	–	–

表5：*VoxEmo* 零样本 + 监督微调结果。*Q2A* =*Qwen2-Audio*、*AF3* = *Audio Flamingo 3*。提示词缩写：*T* = 自动语音识别转录文本，*A* = 声学描述，*R* = 推理。 红色粗体：每个数据集上 *Q2A* 的最佳提示词；红色粗体：每个数据集上 *AF3* 的最佳提示词。红色粗体：每个数据集上 *Q2A* 监督微调 / *AF3* 监督微调 / *EmoBox* 中的最佳分数*[13]*（按宏平均*F1*计）。第 *2/2* 部分。

设置	UA(%)	WA(%)	F1(%)	UA(%)	WA(%)	F1分数(%)	UA(%)	WA(%)	F1分数(%)	UA(%)	加权准确率（%）	F1分数（%）
	Oreau（法语，7）			PAVOQUE（德语，5）			波兰语（波兰语，3）			RAVDESS（英语，8）		
Q2A 直接	46.22	47.12	39.85	**47.56**	**62.77**	**46.09**	34.89	34.89	21.97	**72.4**	**72.43**	**67.55**
Q2A +T	39.84	41.35	37.53	27.72	53	29.07	34	34	27.91	49.15	45.76	44.54
Q2A +A	33.12	35.58	25.72	45.23	67.52	45.27	**34.67**	**34.67**	**30.14**	68.29	66.39	62.73
Q2A +T+A	41.89	44.23	41.82	31.14	59.61	33.85	35.11	35.11	25.8	49.8	47.01	48.57
Q2A +T+A+R	**45.52**	**48.08**	**46.17**	34.54	60.26	36.3	34.44	34.44	27.34	45.12	44.93	44
AF3 直接	33.62	35.58	35.49	**38.3**	**67.3**	**41.11**	34.67	34.67	19.25	**47.85**	**45.42**	**44.66**
AF3 +T	**36.84**	**39.42**	**37.12**	25.58	55.51	27.94	33.11	33.11	19.59	34.9	30.56	32.92
AF3 +A	30.75	32.69	30.24	31.32	64.46	32.98	34	34	18.33	42.12	38.82	40.17
AF3 +T+A	28.17	30.77	24.81	23.46	57.97	22.97	32.44	32.44	19.45	35.03	30.9	34.96
AF3 +T+A+R	32.49	34.62	32.55	20	46.56	22.43	**36.22**	**36.22**	**29.01**	33.46	30.14	34.57
Q2A 监督微调	50.34	51.92	50.85	**89.84**	**94.21**	**89.95**	44.67	44.67	37.02	**87.76**	**90.83**	**87.81**
AF3 监督微调	43.83	46.15	46.13	72.06	84.77	76.31	36	36	28.6	34.44	35.7	34.77
EmoBox[13]	**84.48**	**84.79**	**85.01**	87.73	93.4	88.43	**83.27**	**83.27**	**82.77**	75.32	75.87	75.19
设置	RESD（俄语，7）			SAVEE（英语，7）			ShEMO（波斯语，6）			SUBESCO（孟加拉语，7）		
Q2A 直接	22.38	22.47	16.74	54.41	56.04	51.73	47.3	49.79	43.56	30.39	30.39	24.84
Q2A +T	18.54	18.35	17.04	28.33	31.04	27.1	26.23	30.98	24.88	17.69	17.69	15.13
Q2A +A	23.72	22.1	19.22	**58.33**	**61.88**	**58.51**	**69.85**	**71.57**	**59.43**	**37.23**	**37.23**	**32.39**
Q2A +T+A	20.02	20.23	18.15	42.38	47.5	43.47	33.91	43.14	32.01	25.96	25.96	22.12
Q2A +T+A+R	**23.23**	**23.22**	**22.17**	45.48	50.62	45.65	41.9	46.96	36.44	27.56	27.56	24.07
AF3 直接	**22.8**	**23.22**	**17.14**	**53.1**	**55.42**	**54.57**	52.46	60.96	48.53	**24.73**	**24.73**	**19.18**
AF3 +T	14.71	14.61	11.01	34.64	42.5	34.71	30.7	47.1	33.77	20.57	20.57	14.24
AF3 +A	21.98	22.47	15.76	46.67	52.92	48.44	**49.06**	**64.64**	**50.31**	24.51	24.51	17.77
AF3 +T+A	13.89	13.86	7.87	35.12	42.29	37.06	24.4	40.03	27.12	19.93	19.93	14.12
AF3 +T+A+R	12.02	11.98	6.66	30.83	37.29	32.15	24.26	33.95	26.36	13.6	13.6	13.04
Q2A 监督微调	28.88	30.71	27.18	64.64	65	62.3	**87.16**	**95.19**	**88.28**	**76.29**	**76.29**	**76.16**
AF3 监督微调	20.6	20.97	19.92	33.34	37.71	34.57	57.15	75.95	60.62	31.36	31.36	31.46
EmoBox[13]	**55.87**	**56.47**	**55.82**	**75.65**	**78.25**	**78.38**	80.23	89.55	82.94	73.05	73.05	72.94
设置	TESS（英语，7）			Thai-ser（泰语，5）			TurEV-DB（tr，4）			URDU（ur，4）		
Q2A 直接	71.86	71.86	73.66	**29.21**	**29.17**	**25.49**	**28.11**	**28.93**	**23.83**	38	38	35.52
Q2A +T	36.86	36.86	36.02	17.09	16.45	13.62	21.62	21.76	20.44	22	22	19.05
Q2A +A	**77**	**77**	**73.89**	25.42	24.59	21.56	26.79	27.78	21.39	**36**	**36**	**35.97**
Q2A +T+A	26.86	26.86	24.92	20.12	19.09	14.03	22.77	21.76	15.14	27	27	21.4
Q2A +T+A+R	58.71	58.71	58.44	21.12	20.09	15.18	22.45	21.53	14.14	24	24	19.78
AF3 直接	**59.14**	**59.14**	**61.36**	**30.8**	**29.93**	**23.54**	28.67	31.25	19.43	38	38	32.2
AF3 +T	51.86	51.86	57	23.79	22.11	15.81	23.48	23.15	18.88	29	29	20.93
AF3 +A	55.57	55.57	59.58	28.35	27.78	20.39	**28.36**	**28.47**	**22**	**42**	**42**	**36.85**
AF3 +T+A	53.43	53.43	58.55	22.73	21.47	14.31	22.15	21.07	11.78	34	34	24.15
AF3 +T+A+R	47.14	47.14	52.88	19.92	18.75	12.46	22.04	21.07	12.01	36	36	25
Q2A 监督微调	99.86	99.86	99.86	**72.42**	**72.79**	**72.3**	51.88	53.24	51.36	41	41	35.46
AF3 监督微调	82.71	82.71	83.19	51.95	52.45	51.7	42.49	43.75	42.52	41	41	33.75
EmoBox[13]	**99.96**	**99.96**	**99.96**	–	–	–	**81.32**	**81.58**	**81.31**	**88.41**	**88.41**	**88.4**

在 EmoDB (−23.0)、SAVEE (−20.0) 和 JL-Corpus (−12.0) 上。AF3 在 30 个可比较的数据集上均未超过 EmoBox 基准，其中 CREMA-D (−1.4) 的差距最小。这些结果基于单一 LoRA 配置 (r=8)；AF3 可能会受益于采用不同的超参数，我们将在第 5 节回到这一点。

总体而言，监督微调大幅缩小了语音大语言模型与传统监督基线之间的差距，Q2A 在 30 个可比较的数据集中有 15 个达到或超过了基线水平。剩余差距主要集中在小规模行为语料库中，并且很大程度上取决于基础模型的选择；Q2A 的优势从零样本时在 35 个数据集中领先 26 个，扩大到监督微调后的 36 个数据集中领先 35 个，便体现了这一点。

4.3. 软标签评估

表6全面评估了五个提供多标注者分布的语料库上的零样本软标签表现。根据图1中观察到的提示敏感性，我们选取了三种具有代表性的配置进行详细比较：基线**直接**提示、声学描述提示（**+A**），以及我们提出的**集成** 策略。

4.3.1. 提示词的影响与零样本不稳定性

评估生成式语音大语言模型的一项关键挑战在于推理时随机性，以及严格遵循结构化输出格式。如图1所示，要求模型执行中间推理任务往往会影响其生成有效 JSON 预测的能力。有趣的是，这种性能下降高度依赖于模型。

对于 Q2A，提示复杂度的增加会导致平均解析失败率（红线）灾难性飙升——在 **+A**提示词下达到峰值 93.2%——直接使平均宏 F1 分数骤降至接近零。相较之下，AF3 对 **+A** 提示词的处理相对较好，取得了最高的单提示词 F1 分数（31.4%），且解析失败率较低（10.9%）。尽管如此，当面对组合推理提示（例如，**+T+A+R**）时，两个模型都会出现严重的格式错误。这表明，依赖单一提示词本质上十分脆弱，且在不同 LLM 之间表现不一致。

4.3.2. 通过提示集成缓解不稳定性

为缓解这种零样本不稳定性，我们按照 [61] 对提出的提示词集成策略（**集成**）进行评估。该集成方法通过统一的回退机制汇总所有五种提示变体的离散预测，以处理解析错误

表6： 零样本软标签评估结果。我们比较了单次直接提示（直接）、 +A 提示与所提出的提示集成方法（集成）。性能分为硬决策指标（宏平均 *F1*、*Micro-F1* 和 *Top-1*准确率，单位为%）以及分布感知评估，包括库尔贝克-莱布勒散度（*KLD*）、詹森-香农散度（*JSD*）、总变差距离（*TVD*）、余弦相似度（*SIM* ×100）和均方误差（*MSE*）。红色粗体表示每个语料库中性能更优的设置。↑（↓）表示数值越高（越低）越好。

数据集	模型	提示	硬决策评估			分布感知评估				
			Ma-F1↑	Mi-F1↑	Acc↑	KLD↓	JSD↓	TVD↓	SIM↑	MSE↓
BIIC-Podcast（繁体中文，8）	Q2A	直接	28.14	42.16	29.57	5.473	0.363	0.657	49.61	0.074
		+A	4.79	9.38	9.67	**2.440**	0.402	0.749	45.78	0.073
		集成	**31.12**	**51.14**	**42.95**	7.257	**0.309**	**0.559**	**56.49**	**0.070**
	AF3	直接	15.57	25.70	22.95	4.767	0.374	0.680	46.97	0.083
		+A	16.88	42.70	38.11	10.391	0.358	0.616	48.96	0.101
		集成	20.98	45.81	41.17	10.172	0.347	0.595	50.11	0.097
CREMA-D（英语，6）	Q2A	直接	**61.88**	65.34	49.03	4.112	0.196	0.443	69.43	0.054
		+A	1.99	1.76	16.43	**1.092**	0.253	0.554	60.12	0.058
		集成	61.00	**68.25**	**59.16**	5.482	**0.175**	**0.389**	**75.81**	**0.045**
	AF3	直接	40.57	41.19	39.22	6.372	0.275	0.557	59.71	0.088
		+A	48.59	54.14	53.34	6.024	0.239	0.501	65.67	0.073
		集成	54.06	63.98	57.60	6.652	0.209	0.446	71.34	0.064
EmotionTalk（中文，7）	Q2A	直接	33.10	40.17	36.24	5.842	0.377	0.667	47.91	0.105
		+A	8.95	13.36	21.51	**3.191**	0.428	0.770	41.55	0.111
		集成	**36.23**	47.11	47.33	6.607	0.323	0.583	55.13	0.091
	AF3	直接	23.80	31.41	31.99	3.599	0.369	0.661	49.19	0.100
		+A	31.15	**54.39**	51.84	7.352	0.313	0.535	56.32	0.105
		集成	36.09	53.42	**54.07**	5.807	**0.282**	**0.499**	**60.37**	**0.087**
IEMOCAP（英语，4）	Q2A	直接	**62.83**	**62.10**	**60.03**	2.962	**0.234**	0.464	**69.01**	**0.121**
		+A	1.30	1.32	20.53	**1.286**	0.338	0.680	54.53	0.158
		集成	56.00	59.24	54.23	5.369	0.254	0.472	63.66	0.152
	AF3	直接	50.95	55.30	51.89	4.716	0.261	0.465	62.14	0.160
		+A	49.16	53.25	51.15	4.723	0.265	0.480	61.70	0.158
		集成	60.07	61.81	56.85	5.700	0.248	**0.449**	64.80	0.155
MSP-Podcast 2（test1，8）	Q2A	直接	24.43	34.72	18.85	10.371	0.421	0.715	37.14	0.087
		+A	1.75	2.06	19.28	**1.741**	0.361	0.692	49.10	**0.060**
		集成	**33.76**	46.99	36.65	9.959	0.346	0.602	50.52	0.072
	AF3	直接	23.02	33.80	34.35	10.007	0.388	0.681	44.51	0.098
		+A	24.11	36.36	33.33	7.541	0.359	0.656	49.50	0.082
		集成	30.18	**48.61**	**41.62**	10.415	**0.328**	**0.587**	**53.44**	0.078
MSP-Podcast 2（test2，8）	Q2A	直接	24.50	40.16	20.99	7.744	0.361	0.651	46.70	0.068
		+A	1.33	2.17	5.84	**1.500**	0.334	0.649	52.35	**0.053**
		集成	**29.13**	48.04	43.68	9.292	0.313	0.558	57.11	0.059
	AF3	直接	17.98	34.76	39.80	8.283	0.321	0.593	55.39	0.074
		+A	18.80	38.34	41.05	6.589	0.301	0.578	59.59	0.063
		集成	23.38	**48.65**	**55.70**	10.059	**0.284**	**0.521**	**63.11**	0.061

（详见第3.2节）。如表6所示，集成方法有效消除了单个提示词的格式错误，并且始终优于基线**直接** 提示词。它不仅恢复并提升了硬决策评估指标（例如，将 Q2A 在 CREMA-D 上的宏平均 F1 从 61.88% 提高到 68.25%），还显著改善了主观性对齐。在大多数数据集上，**集成** 方法取得了更高的 SIM 和更低的 JSD/TVD。这表明，多样化提示响应所具有的随机性并非缺陷；通过汇总这些响应，可以有效地近似人类情绪感知中固有的不确定性和主观性。

*4.3.3. Qwen2-Audio*与*AudioFlamingo 3*的对比

在大多数参与基准评测的语料库中，Q2A与人类软标签的对齐表现始终优于AF3。这一点在CREMA-D和IEMOCAP数据集上尤为明显，Q2A取得了显著更高的余弦相似度（↑）和更低的均方误差（↓）。这些结果表明，Q2A的底层表征在结构化环境中更能模拟人类标注者的共识。

然而，在 EmotionTalk 和 BIIC-Podcast 等自然场景数据集中，可以观察到性能权衡：尽管 AF3 的准确率较低，但偶尔能取得更低的 KLD 和 JSD。这种差异表明，虽然 Q2A 可能在 top-1 预测中更准确，但其概率质量往往过度集中（过度自信）；相比之下，AF3 的分布更平滑，能够更好地反映跨语言自然语音中固有的不确定性。

表7：最佳零样本语音大语言模型与近期全监督最先进基准在硬标签宏平均*F1*（%）上的比较。

数据集	模型／框架	Ma-F1
BIIC-Podcast（繁体中文，8）	[64]	35.5
	[6]	28.3
	本文（Q2A-Ensemble）	31.1
	本文（AF3-Ensemble）	21.0
CREMA-D（英语，6）	[21]	71.0
	[64]	70.6
	[6]	67.6
	本文（Q2A-Direct）	61.9
	本文（AF3-Ensemble）	54.1

4.3.4. 捕捉情感歧义与主观性

为比较生成式语音大语言模型在更广泛的语音情感识别领域中的表现，表7将表现最佳的零样本配置与近期全监督的最先进（SOTA）基准进行比较，其中包括传统的 Transformer [64], 自监督学习（SSL）表征 [6], 以及分布感知监督模型 [21]。如表7所示，零样本语音大语言模型的表现通常不及以下模型：

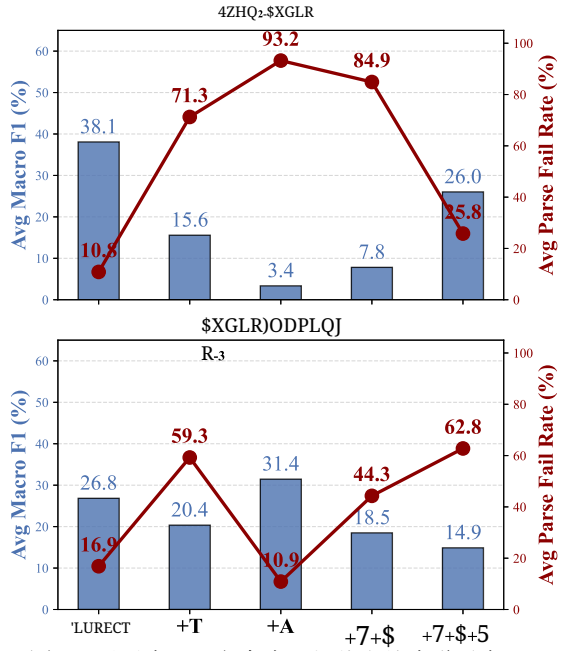

图1： 零样本提示复杂度对性能和格式遵循度的影响。蓝色柱状图（左轴）表示五个评估语料库的平均宏 *F1*，红色曲线（右轴）表示平均解析失败率。随着提示约束增加（从直接分类 *Direct* 到要求生成中间声学描述 *+T*），*Q2A* 和 *AF3* 都难以遵循结构化文本输出格式，导致整体预测性能严重下降。这种零样本随机性使我们有必要提出分布感知提示集成方法。

特定任务监督模型在硬标签宏平均F1方面的表现。然而，这种性能差距凸显的是建模范式上的根本差异，而非单纯的能力不足。监督基线模型通过交叉熵进行显式优化，将主观标注归并为硬标签，从而最大限度地提高 F1 等分类指标。相比之下，零样本生成式语音大语言模型保留了更广泛、未经调优的概率质量。

这引出了本文的一项关键贡献：对情感歧义的量化。在 MSP-Podcast 系列等以评分者间高度变异为特征的自然场景数据集中，Q2A 与 AF3 相对较低的 TVD 和 JSD 分数（见表6）表明，语音大语言模型具备建模情感主观性的强大潜在能力。与会丢弃人类感知差异的传统闭集分类器不同，生成式语音大语言模型会在其输出逻辑值中自然表征相互竞争的情绪状态。这些模型无需任何特定数据集微调，即可对人类标签分布保持较高的 SIM，这证明它们在开发尊重多位听者多样感知的情感感知系统方面具有非凡潜力。

4.4. 跨语料库评估

表 8 报告了两个模型在使用各数据集原生标签集时的跨域迁移结果。对于 Q2A，在不匹配的英语源数据集上进行微调，其效果在 12 个源 - 目标数据集对中的 11 个上优于最佳零样本提示。MELD 是效果最好的源数据集，在三个目标数据集上的宏 F1 提升平均达到 17.7。迁移至 IEMOCAP 时，分数达到 81.6；

表 8：跨场景迁移（宏平均 *F1*）。每个单元格显示在源域（行）上微调并在目标域（列）上评估时的分数，括号内为相对最佳零样本的 Δ。对角线（粗体）：域内监督微调。*MSP-Podcast* 使用 *test1*。

源\目标 CREMA-D	IEMOCAP	MELD	MSP-Podcast 2
Qwen2-Audio			
CREMA-D	**91.1**(+21.7)	79.9 (+20.5)	30.7 (+2.6)	25.8 (+5.5)
IEMOCAP	76.8 (+7.4)	**82.4** (+23.0)	24.7 (-3.4)	31.5 (+11.1)
MELD	86.6 (+17.1)	81.6 (+22.2)	**39.3** (+11.2)	34.1 (+13.7)
MSP-Podcast 2	78.6 (+9.1)	68.1 (+8.7)	30.8 (+2.8)	**29.2** (+8.8)
AudioFlamingo-3				
CREMA-D	**75.2** (+15.4)	65.1 (+1.1)	21.6 (-8.2)	22.9 (-5.7)
IEMOCAP	63.6 (+3.8)	**65.2** (+1.2)	20.3 (-9.5)	19.2 (-9.5)
MELD	63.3 (+3.5)	62.0 (-2.0)	**25.7** (-4.1)	23.5 (-5.2)
MSP-Podcast 2	36.6 (-23.2)	45.4 (-18.6)	18.9 (-10.9)	**19.1** (-9.6)

迁移至 CREMA-D 时，分数为 86.6。另一个真实场景语料库 MSP-Podcast 也提升了全部三个目标数据集上的表现，但平均提升幅度明显较小。相反，MELD 也是最难迁移的目标数据集：表中唯一一次性能下降出现在 IEMOCAP→MELD (−3.4) 这一迁移中，因为 IEMOCAP 的 4 个类别仅涵盖 MELD 的 7 个类别中的 4 个，缺少厌恶、恐惧和惊讶。

在相同的 LoRA 配置下，AF3 仅在 12 组配对中的 3 组超过零样本表现，且这些配对的目标均为表演式语料库，提升幅度有限。所有 6 组迁移到真实场景目标（ MELD 、 MSP-Podcast ）的表现均有所下降，这与 AF3 在这两个语料库上的域内监督微调表现已经低于零样本结果一致。MSP-Podcast 是一个特别有害的源数据集，造成了表中降幅最大的三项结果。与第4.2节中的情况一样，这些结果可能部分反映了 LoRA 超参数不匹配，而非模型存在根本性局限。

5. 结论、局限性与未来工作

本文介绍了 VoxEmo，这是一个评估工具包和基准，用于规范基于语音大语言模型的语音情感识别在 15 种语言的 35 个语料库上的推理时协议。零样本性能对提示设计高度敏感，而语料库构建中的表演式—自然式划分会系统性地影响提示效果。监督微调缩小了与传统基线的差距，但在大多数情况下并未消除差距，其效果取决于数据集规模和基础模型的选择。尽管如此，即使没有针对任务的训练，零样本输出也能捕捉与人工标注分布相符的情感歧义；生成式接口还支持在不匹配的标签集之间进行跨域迁移，不过这两项优势在不同模型间差异显著。

需要指出若干局限性。本基准仅评估了两个采用相同音频编码器（Whisper-large-v3）且参数规模相同的模型；观察到的模式能否推广至规模更大的模型，或采用不同音频前端（包括近期基于分词器的架构）的模型，仍有待研究。目前，这类基于分词器的模型无法在相同的推理环境中运行，因此限制了直接比较。全文采用单一的 LoRA 配置，而 AF3 的监督微调收益有限，部分原因可能是超参数不匹配。软标签评估仅限于提供逐标注者元数据的五个语料库。此外，本文聚焦于语料库级聚合指标，未考察类别不平衡、逐类别性能或说话人效应等数据集内因素。未来工作将把基准扩展至架构多样的模型，研究监督微调超参数敏感性，并纳入更细致的逐数据集分析。

6. 生成式人工智能使用披露

生成式人工智能工具仅用于改进语法、清晰度和可读性。作者对本文呈现的所有技术内容和结果负全部责任。这些工具未用于生成科学创意、分析或研究发现。

7. 参考文献

[1] B. W. Schuller，“语音情感识别：二十年发展概览、基准与当前趋势”，《美国计算机协会通讯》，第61卷，第5期，第90–99页，2018年4月。[2] C.-C. Lee、T. Chaspari、E. M. Provost 和 S. S. Narayanan，“情绪与语音的工程学视角：从分析和预测模型到负责任、以人为本的应用”，《*IEEE*学报》，第111卷，第10期，第1142–1158页，2023年。[3] C. M. Lee 和 S. Narayanan，“迈向口语对话中的情绪检测”，《*IEEE*语音与音频处理汇刊》，第13卷，第2期，第293–303页，2005年。[4] P. Gupta 和 N. Rajput，“用于呼叫中心监控的双流情绪识别”，发表于*Interspeech 2007*，2007年，第2241–2244页。[5] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao、J. Wu、L. Zhou、S. Ren、Y. Qian、Y. Qian、J. Wu、M. Zeng、X. Yu 和 F. Wei，“WavLM：面向全栈语音处理的大规模自监督预训练”，《*IEEE*信号处理精选主题期刊》，第16卷，第6期，第1505–1518页，2022年。[6] H. Wu、H.-C. Chou、K.-W. Chang、L. Goncalves、J. Du、J.-S. R. Jang、C.-C. Lee 和 H.-Y. Lee，“Open-Emotion：面向语音情感识别系统的可复现 EMO-Superb”，发表于*2024*年*IEEE*口语技术研讨会（*SLT*），2024年，第510–517页。[7] J. Bellver、I. Mart´ ín-Fern´andez、J. M. Bravo-Pacheco、S. Es- teban、F. Fern´ andez-Mart´ınez 和 L. F. D’Haro，“用于语音情感识别的多模态音频语言模型”，发表于语音与语言识别研讨会（*Odyssey 2024*），2024年，第288–295页。[8] Y. Chu、J. Xu、Q. Yang、H. Wei、X. Wei、Z. Guo、Y. Leng、Y. Lv、J. He、J. Lin、C. Zhou 和 J. Zhou，“Qwen2-Audio 技术报告”，2024年7月。[9] J. Mai、X. Xing、W. Chen、Y. Fang 和 X. Xu，“AA-SLLM：用于语音情感识别的声学增强型语音大语言模型”，发表于*Interspeech2025*。ISCA，2025年8月，第4328–4332页。[10] W. Cui、D. Yu、X. Jiao、Z. Meng、G. Zhang、Q. Wang、S. Y. Guo 和 I. King，“语音语言模型的最新进展：综述”，发表于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），W. Che、J. Nabende、E. Shutova 和 M. T. Pilehvar 编。奥地利维也纳：计算语言学协会，2025年7月，第13 943–13 970页。[11] S.-w. Yang、P.-H. Chi、Y.-S. Chuang、C.-I. J. Lai、K. Lakhotia、Y. Y. Lin、A. T. Liu、J. Shi、X. Chang、G.-T. Lin、T.-H. Huang、W.-C. Tseng、K.-t. Lee、D.-R. Liu、Z. Huang、S. Dong、S.-W. Li、S. Watanabe、A. Mohamed 和 H.-y. Lee，“SUPERB：语音处理通用性能基准”，发表于*Interspeech 2021*。ISCA，2021年8月，第1194–1198页。[12] L. Pepino、P. Riera 和 L. Ferrer，“使用 wav2vec 2.0 嵌入进行语音情感识别”，发表于*Interspeech 2021*。ISCA，2021年8月，第3400–3404页。[13] Z. Ma、M. Chen、H. Zhang、Z. Zheng、W. Chen、X. Li、J. Ye、X. Chen 和 T. Hain，“EmoBox：多语言多语料库语音情感识别工具包与基准”，2024年6月。[14] J. Murzaku 和 O. Rambow。《OmniVox：使用 Omni-LLM 进行零样本情绪识别》。 [在线]。可访问：http://arxiv.org/abs/2503.21480

[15] M. M. Amin 和 B. W. Schuller，“论 ChatGPT 在情感计算中的提示敏感性”，载于*2024*年第*12*届情感计算与智能交互国际会议（*ACII*）。IEEE，2024年，第 203–209 页。[16] P. Larrouy-Maestri、D. Poeppel 和 M. D. Pell，“情感韵律之声：近三十年的研究与未来方向”，《心理科学展望》，第 20 卷，第 4 期，第 623–638 页，2025年7月。[17] H.-C. Chou 和 C.-C. Lee，“重新审视语音情感识别中的建模与评估方法：考量标注者的主观性与情绪的歧义性”，2025年。[在线]。可访问：https://arxiv.org/abs/2510.05934[18] R. Lotfian 和 C. Busso，“通过从现有播客录音中检索情感语音，构建自然情境下情绪均衡的语音语料库”，《IEEE情感计算汇刊》，第 10 卷，第 4 期，第 471–483 页，2019年10月。[19] C. Busso、R. Lotfian、K. Sridhar、A. N. Salman、W.-C. Lin、L. Goncalves、S. Parthasarathy、A. R. Naini、S.-G. Leem、L. Martinez-Lucas、H.-C. Chou 和 P. Mote，“MSP播客语料库”，2025年9月。[20] Carlos Busso 和 S. S. Narayanan，“情绪的表达与感知：比较对自身与他人的评估”，载于*Interspeech2008*。ISCA：ISCA，2008年9月，第 257–260 页。[21] H.-C. Chou、L. Goncalves、S.-G. Leem、A. N. Salman、C.-C. Lee 和 C. Busso，“少数观点很重要：采用全包容聚合规则，以人工主观标注评估语音情感分类器”，《IEEE情感计算汇刊》，第 16 卷，第 1 期，第 41–55 页，2025年1月。[22] H.-C. Chou、H. Wu、L. Goncalves、S.-G. Leem、A. Salman、C. Busso、H.-Y. Lee 和 C.-C. Lee，“运用全包容聚合规则，在评估多标签语音情感识别系统时拥抱歧义性与主观性”，载于*2024*年*IEEE*口语技术研讨会（*SLT*），2024年12月，第 502–509 页。[23] R. Milner、M. A. Jalal、R. W. Ng 和 T. Hain，“语音情感识别的跨语料库研究”，载于*2019*年*IEEE*自动语音识别与理解研讨会（*ASRU*）。IEEE，2019年，第 304–311 页。[24] O. C. Phukan、M. M. Akhtar、S. R. Behera、P. Bhagath、P. B. Reddy、A. B. Buduru等，“重新思考跨语料库语音情感识别基准测试：副语言预训练表征是否足够？”，载于*2025*年亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*）。IEEE，2025年，第 1022–1027 页。[25] H. Sun、X. Wang、J. Zhao、S. Zhao、J. Zhou、H. Wang、J. He、A. Kong、X. Yang、Y. Wang、Y. Lin 和 Y. Qin，“EmotionTalk：一个具有丰富标注的交互式中文多模态情感数据集”，2025年6月。[26] S. G. Upadhyay、W.-S. Chien、B.-H. Su、L. Goncalves、Y.-T. Wu、A. N. Salman、C. Busso 和 C.-C. Lee，“面向大规模自然情境情感语音语料库采集的智能基础设施”，载于*2023*年第*11*届情感计算与智能交互国际会议（*ACII*），2023年，第 1–8 页。[27] J. Zhao、T. Zhang、J. Hu、Y. Liu、Q. Jin、X. Wang 和 H. Li，“M3ED：多模态多场景多标签情感对话数据库”，载于计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），S. Muresan、P. Nakov 和 A. Villavicencio 编。爱尔兰都柏林：计算语言学协会，2022年5月，第 5699–5710 页。 [在线]。可访问：https://aclanthology.org/2022.acl-long.391/[28] T. Landry Dejoli、Q. He、H. Yan 和 Y. Li，“ASVP-ESD：一个使用语音与非语音话语进行情感识别的数据集及其基准”，载于*GSJ*论文集，2020年。[29] S. Poria、D. Hazarika、N. Majumder、G. Naik、E. Cambria 和 R. Mihalcea，“MELD：一个用于对话情感识别的多模态多方数据集”，2019年6月。

[30] S. Latif、A. Qayyum、M. Usman 和 J. Qadir，“跨语言语音情感识别：URDU 与西方语言的比较”，载于*FIT*论文集，2018年。

[31] F. Burkhardt、A. Paeschke、M. Rolfes、W. F. Sendlmeier 和 B. Weiss，《德语情感语音数据库》，载于*Inter- speech2005*。ISCA，2005年9月，第1517–1520页。

[32] F. Burkhardt、O. Schr¨ufer、U. Reichel、H. Wierstorf、A.Derington、F. Eyben 和 B. W. Schuller，《EmoDB 2.0：非黑非白、而是灰色世界中的情感语音数据库》，载于*Interspeech2025*。ISCA，2025年8月，第4488–4492页。

[33] J. Wongpithayadisai、C. Chaksangchaichot、S. Sangnark、P. Prakrankamanant、K. Gangwanpongpun、S. Boonpunmongkol、P. Milindasuta、D. Na-Pombejra、S. Nutanong 和 E. Chuang-suwanich，《泰语语音情感识别（THAI-SER）语料库》，2025年7月。

[34] I. Christop，《nEMO：波兰语情感语音数据集》，载于*2024*年计算语言学、语言资源与评估联合国际会议（*LREC-COLING 2024*）论文集，N. Calzolari、M.-Y. Kan、V. Hoste、A. Lenci、S. Sakti 和 N. Xue 编。意大利都灵：ELRA 和 ICCL，2024年5月，第 12 111–12 116 页。 [在线]。网址：https://aclanthology.org/2024.lrec-main.1059/

[35] E. A. Retta、E. Almekhlafi、R. Sutcliffe、M. Mhamed、H. Ali 和 J. Feng，《一项新的阿姆哈拉语语音情绪数据集及分类基准》，载于*TALLIP* 论文集，2023年。

[36] K. A. Noriy、X. Yang 和 J. J. Zhang，《EMNS /Imz/语料库：用于游戏、电视和图像小说叙事讲故事的情感单说话人数据集》，2023年5月。

[37] F. 卡塔尼亚，《使用 Wav2Vec 2 进行意大利语语音情感识别》，载于 *Authorea*预印本，2023年。

[38] I. Lubenets、N. Davidchuk 和 A. Amentes，“Aniemore。”[在线]。网址：https://github.com/aniemore/Aniemore[39] K. Zhou、B. Sisman、R. Liu 和 H. Li，“使用新型情感语音数据集进行已见与未见情感风格的语音转换”，载于*Proc. ICASSP*，2021年。[40] M. M. Duville、L. M. Alonso-Valerdi 和 D. I. Ibarra-Zarate，“墨西哥情感语音数据库（MESD）：基于机器学习的构建与评估”，载于*2021*年第*43*届*IEEE*医学与生物学工程学会国际年会，2021年11月，第1644–1647页。[41] K. Wang、Q. Wu、L. Song、Z. Yang、W. Wu、C. Qian、R. He、Y. Qiao 和 C. C. Loy，“MEAD：用于情感说话人脸生成的大规模视听数据集”，载于计算机视觉——*ECCV 2020*，A. Vedaldi、H. Bischof、T. Brox 和 J.-M. Frahm 编。Cham：施普林格国际出版，2020年，第12366卷，第700–717页。[42] L. KERKENI、C. CLEDER、S.-R. Youssef 和 K. RAOOF，“法语情感语音数据库-or´eau”，2020年。[43] M. Miesikowska 和 D. Swisulski，“波兰语语音录音中的情绪”，https://doi.org/10.34808/h46c-hb44，2020年。[44] S. F. Canpolat、Z. Ormano˘glu 和 D. Zeyrek，“土耳其语情感语音数据库（TurEV-DB）”，载于第一届口语低资源语言技术联合研讨会（*SLTU*）与低资源语言协作与计算（*CCURL*）论文集，D. Beermann、L. Be- sacier、S. Sakti 和 C. Soria 编。法国马赛：欧洲语言资源协会，2020年5月，第368–375页。[45] O. Mohamad Nezami、P. Jamshid Lou 和 M. Karami，“ShEMO：一个大规模、经过验证的波斯语语音情感检测数据库”，《语言资源与评测》，第53卷，第1期，第1–16页，2019年3月。[46] S. Sultana、M. S. Rahman、M. R. Selim 和 M. Z. Iqbal，“SUST孟加拉语情感语音语料库（SUBESCO）：一个孟加拉语纯音频情感语音语料库”，*PLOS ONE*，第16卷，第4期，文章编号e0250173，2021年4月。

[47] N. Vrysas、R. Kotsakis、A. Liatsou、C. Dimoulas 和 G. Kalliris，“表演互动中的语音情感识别”，音频工程学会期刊，第66卷，第6期，第457–467页，2018年6月。[48] P. Gournay、O. Lahaie 和 R. Lefebvre，“加拿大法语情感语音数据集”，载于第九届*ACM*多媒体系统会议论文集。荷兰阿姆斯特丹：ACM，2018年6月，第399–402页。[49] A. Adigwe、N. Tits、K. E. Haddad、S. Ostadabbas 和 T. Dutoit，“情感语音数据库：迈向语音生成系统中情绪维度的控制”，2018年6月。[50] J. James、L. Tian 和 C. Inez Watson，“一个用于人机交互应用的开源情感语音语料库”，载于*Interspeech 2018*。ISCA，2018年9月，第2768–2772页。[51] S. R. Livingstone 和 F. A. Russo，“瑞尔森情感语音与歌曲视听数据库（RAVDESS）：北美英语中动态、多模态的面部和声音表达集”，*PLOS ONE*，第13卷，第5期，文章编号e0196391，2018年5月。[52] H. Cao、D. G. Cooper、M. K. Keutmann、R. C. Gur、A. Nenkova 和 R. Verma，“CREMA-D：众包情感多模态演员数据集”，*IEEE*情感计算汇刊，第5卷，第4期，第377–390页，2014年10月。[53] G. Costantini、I. Iadarola、A. Paoloni 和 M. Todisco，“EMOVO语料库：意大利语情感语音数据库”，2014年。[54] P. Jackson 和 S. Haq，“萨里视听表达情感（SAVEE）数据库。”萨里大学，2014年。[55] I. Steiner、M. Schr¨oder 和 A. Klepp，“PAVOQUE语料库：表现性语音分析与合成资源”，载于语音学与音韵学论文集，2013年。[56] K. Dupuis 和 M. K. Pichora-Fuller，“多伦多情感语音集（TESS）——年轻说话者高兴情绪。”多伦多大学，2010年。[57] J. Tao、F. Liu、M. Zhang 和 H. Jia，“普通话文本转语音语料库的设计”，载于*Blizzard*挑战赛研讨会，2008年。[58] O. Martin、I. Kotsia、B. Macq 和 I. Pitas，“eNTERFACE’ 05视听情感数据库”，载于第*22*届数据工程国际会议研讨会（*ICDEW’06*），2006年4月，第8–8页。[59] A. S. Cowen 和 D. Keltner，“语义空间理论：情感的计算方法”，认知科学趋势，第25卷，第2期，第124—136页，2021年2月。 [在线]。网址：https://doi.org/10.1016/j.tics.2020.11.004[60] A. Goel、S. Ghosh、J. Kim、S. Kumar、Z. Kong、S.-g. Lee、C.-H. H. Yang、R. Duraiswami、D. Manocha、R. Valle 和 B. Catanzaro，“Audio Flamingo 3：利用完全开放的大型音频语言模型推进音频智能”，2025年7月。[61] W. Zhang、H. Jin、S. Wang、Z. Wei 和 T. Dang，“扩展歧义：利用音频语言模型增强语音情感识别中的人工标注”，载于*ICASSP 2026——2026*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2026年，第1–5页。[62] B.-H. Su、H.-Y. Shih、J. Tian、J. Shi、C.-C. Lee、C. Busso 和 S. Watanabe，“超越多数投票的推理：用于语音情感识别的可解释语音语言模型框架”，*arXiv*预印本 *arXiv:2509.24187*，2025年。[63]E. J. Hu、yelong shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang 和 W. Chen，“LoRA：大语言模型的低秩适配”，载于国际学习表征会议，2022年。 [在线]。网址：https://openreview.net/forum?id=nZeVKeeFYf9[64] J. Wagner、A. Triantafyllopoulos、H. Wierstorf、M. Schmitt、F. Burkhardt、F. Eyben 和 B. W. Schuller，“语音情感识别中的变换器时代曙光：弥合效价差距”，*IEEE*模式分析与机器智能汇刊，第45卷，第9期，第10 745–10 759页，2023年。