

超越分类：使用多任务音频大型语言模型进行语音情感推理

Wenyu Zhang¹, Yingxu He^{1*}, Geyu Lin^{1*}, Zhuohan Liu^{1*}, Shuo Sun^{1*}, Bin Wang^{1*},
Xunlong Zou^{1*}, Jeremy H. M. Wong¹, Qiongqiong Wang¹, Hardik B. Sailor¹,
Nancy F. Chen^{1,2}, Ai Ti Aw¹

¹Institute for Infocomm Research (I²R), Agency for Science, Technology and Research (A*STAR)

²Centre for Frontier AI Research (CFAR), Agency for Science, Technology and Research (A*STAR)
Singapore

Abstract

音频大型语言模型 (AudioLLMs) 在语音识别和翻译等语义任务中取得了优异的成绩，但在情感等副语言线索的建模上仍然有限。现有的方法通常将情感理解视为一个分类问题，几乎没有提供对预测背后原因的深入见解。在这项工作中，我们探讨了情感推理，这是一种利用 AudioLLMs 的生成能力，通过生成语义对齐、证据支撑的解释来增强情感识别的策略。为了在多任务的 AudioLLMs 中支持这一点，我们引入了一个统一框架，结合了推理增强的数据监督、双编码器架构和任务交替训练。该方法使得 AudioLLMs 能够在有效学习不同任务的同时融入情感推理。在 IEMOCAP 和 MELD 上的实验表明，我们的方法不仅提高了情感预测的准确性，还增强了生成响应的连贯性和证据基础。

1 介绍

音频大语言模型 (AudioLLMs) (MERaLiON Team, 2024; Tang et al., 2024; Chu et al., 2023; Hu et al., 2023; Das et al., 2024; D’efosse et al., 2024) 的最新进展在口语语言理解方面取得了显著进步，尤其是在自动语音识别 (ASR)、语音翻译 (ST) 和口语问答 (SQA) 等关注语义内容的任务中。这些模型通常依赖大规模的音频文本对齐来将语音输入与文本输出对齐 (Ji et al., 2024; Held et al., 2024)。然而，目前的 AudioLLMs 在建模副语言信息 (如情感) 方面存在局限，而这些信息对于需要情感感知或同理心机器行为的应用至关重要 (Wang et al., 2024a; Sakshi et al., 2025; Ao et al., 2024)。

传统的语音情感识别方法主要集中在类别分类 (例如，预测说话者是生气还是悲伤) (Ma et al., 2024; Fu et al., 2025; Zhao et al., 2025)。虽然这种方法在高层次情感检测中很有效，但在情感为何被表达的问题上提供的解释性或推理却很少。

在这项工作中，我们利用 AudioLLMs 的生成能力引入推理 (Ma et al., 2025; Xie et al., 2025)，以改善情感识别。我们不是将情感理解仅视为一个纯粹的判别任务，而是引导模型生成具有基础和语义对齐的解释，反映出说了什么 (语义内容) 以及如何说的 (副语言线索)。我们在图 1 中将情感识别输出分为三种类型：(1) 仅标签：直接分类 (例如，“说话者感到愤怒”)，没有解释或依据；(2) 解释性推理：通过释义后的意图或推测的状态进行解释 (例如，由于反复失败而表达挫折)；(3) 证据基础推理：最理想的形式，结合情感标签和引用的话语 (例如，“我不会再重新开始”)，并对其进行解释以证明情感状态的合理性。

为此，我们提出了一种新的多任务 AudioLLM 框架，结合跨数据、架构和训练的多方面策略。为了引导模型的生成能力，我们从转录对齐的数据中构建了增强推理的监督信号，使得模型能够学习生成基于语言和副语言证据的情感解释。我们的模型架构采用双编码器设计，将语义和情感表征分离。我们还提出了一种任务交替训练策略，分别优化语义和情感编码器以满足各自的目标，旨在平衡各个任务的性能。我们的框架在用于情感和情感识别的基准数据集上进行了评估，即 IEMOCAP (Busso et al., 2008) 和 MELD (Poria et al., 2019)，以及 ASR 和 SQA 任务。总结而言，我们的主要贡献为：

- 我们提出了一种增强推理的方法用于语音情感识别，使得 AudioLLMs 能够生成语义对齐、基于证据的解释，从而增强解释性和预测准确性。
- 我们在数据构建 (推理目标创建)、架构 (双编码器设计) 和训练 (任务交替训练) 方面引入了一个具有多方面策略的统一框架，用于多任务 AudioLLMs。
- 我们进行了全面的实验，结果显示我们的方法有效地平衡了不同任务的性能，在对其他任务影响最小的情况下提高了情感预测，并使生成的情感推理具有连贯性和基础性。

* Contributed equally to this work; authors are listed in alphabetical order.

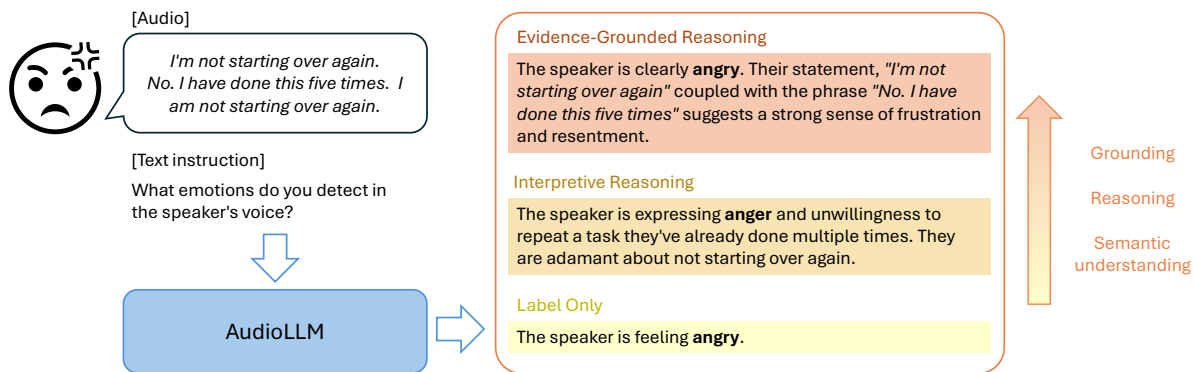


Figure 1: 概述：我们提出的方法利用了 AudioLLM 的生成能力，不仅仅局限于分类，还能产生情感标签以及基于输入语音的语义和副语言内容的解释。

2 相关工作

2.1 音频大型语言模型

多模态大型语言模型 (LLMs)，包括 AudioLLMs (MERaLiON Team, 2024; Tang et al., 2024; Chu et al., 2023; Deshmukh et al., 2023; Hu et al., 2023; Das et al., 2024)，通常采用包括三个核心组件的模块化架构：(1) 一个特定模态的编码器，用于从非文本输入中提取特征，(2) 一个投影或适配器模块，将这些特征映射到与 LLM 的分词器兼容的表示空间中，以及 (3) 一个预训练的 LLM，基于投影的模态标记和自然语言提示生成自由形式的文本响应。例如，Qwen-Audio (Chu et al., 2023) 将 Whisper-large-v2 (Radford et al., 2023) 语音编码器连接到 Qwen-7B (Bai et al., 2023) 语言模型。为了捕捉更丰富的音频表示，一些模型使用双编码器分别建模语义和声学信息。SALMONN (Tang et al., 2024) 将 Whisper-large-v2 和 BEATs (Chen et al., 2023) 与 Vicuna-13B (Chiang et al., 2023) 整合，而 WavLLM (Hu et al., 2023) 则利用 Whisper-large-v2 和 WavLM-base (Chen et al., 2022)，并与 LLaMA-2-chat-7B (Touvron et al., 2023) 接口。

蒸馏方法使用大型语言模型 (LLM) 从语音转录或诸如性别和情感等元数据中生成响应，用于训练音频大型语言模型 (AudioLLMs)。Kang et al. (2024) 使用 LLM 对富有表现力的语音提示生成响应，Wang et al. (2024b) 生成情感感知的文本续篇，Lu et al. (2024a,b) 生成反映写作风格和语气的详细描述。

最近的研究探索了如何使音频大型语言模型能够进行推理。Audio-CoT (Ma et al., 2025) 评估了无训练的思维链提示，其需要音频大型语言模型能够遵循一般指令。Audio-Reasoner (Xie et al., 2025) 使用一个由计划、描述、推理和总结步骤组成的结构化推理框架进行训练。(Li et al., 2025) 提出音频推理器中的复杂推理

过程可能不是必要的，最佳实践仍是一个开放的研究问题。

2.2 音频 LLMs 中的情感识别

最近关于 AudioLLMs 中情感识别的研究探索了多种策略，以增强从语音中进行情感理解的能力 (Bellver et al., 2024)。这些方法利用对话上下文、副语言线索和 ASR 生成的转录文本以提高识别准确率。例如，Sun et al. (2024) 在一个级联管道中采用 ASR 和 LLMs 来转录和分析情感内容，尽管这种管道容易受到错误传播的影响。SECap (Xu et al., 2024) 采用对比和互信息学习来解开语音中的语义和情感表示。Fu et al. (2025) 通过提示 LLMs 根据听者的反应推断情感状态从而建模说话者特征。C\$SER (Zhao et al., 2025) 结合 Whisper 和 Emotion2Vec 编码器以及 Chain-of-Thought 提示，以将情境推理注入到情感分类中。SpeechCueLLM (Wu et al., 2024) 在提示中引入描述性线索，如音量、音高和语速，以丰富 LLM 输入的韵律信息。

与之前主要通过架构或输入级别的增强来提高情感分类准确性的研究相比，我们的方法将范式转向情感推理。我们并不是输出单一的情感标签，而是利用 AudioLLMs 的生成能力，为情感状态生成语义连贯、基于证据的解释。

3 提出的方法

我们提出了一种双编码器多任务 AudioLLM 框架，可以联合建模语音内容和情感推理。我们的架构整合了一个通用语音编码器和一个专门以情感为中心的编码器，随后连接到一个大型语言模型 (LLM)。为了促进丰富的监督，我们引入了从语音转录和情感标签中推导出的推理增强训练目标。此外，我们采用了一种任务交替训练策略，以确保模块化专业化和互补特性的有效融合。

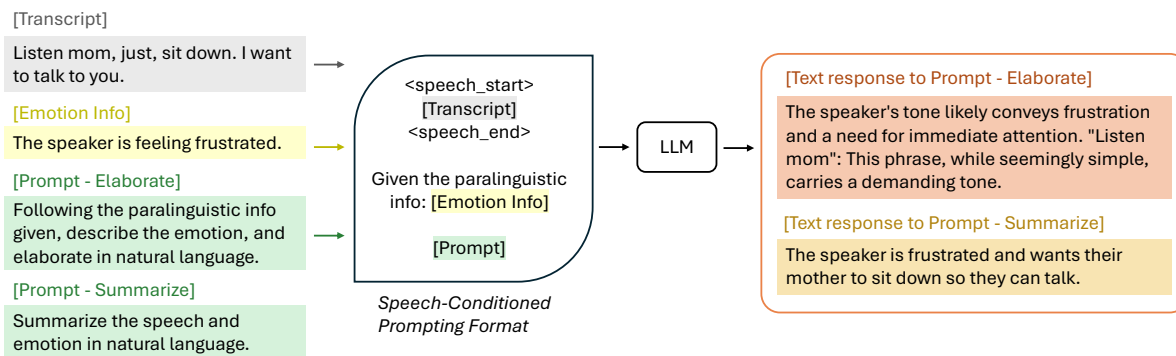


Figure 2: 情感原因提取：我们将转录文本、情感标签和推理提示输入到一个以语音为条件的提示模板中，以引出来自大型语言模型的有根据且语义对齐的情感解释。总结提示鼓励基于隐含背景的解释性推理，而详细提示则鼓励基于转录文本中明确线索的证据支持推理。对于更粗略的情感原因提取，将推理提示中的“情感”一词替换为“情绪（积极、消极、中立）”。

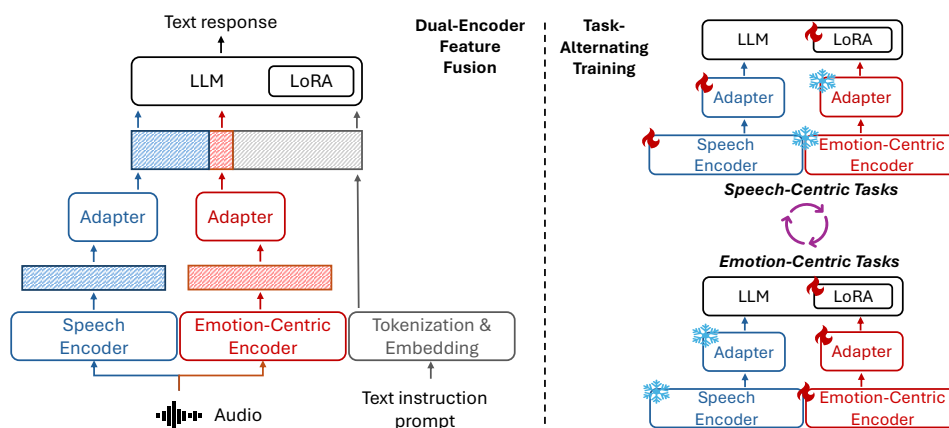


Figure 3: 双编码器特征融合和任务交替训练：我们结合了来自通用语音编码器和情感中心编码器的特征。每个编码器及其适配器通过在以语音为中心和以情感为中心的任务之间交替进行训练，从而实现了具有解耦表示的有效多任务学习。

3.1 情感原因提取

我们引入了增强推理的训练目标，将离散的情感标签与自然语言解释结合。图 2 展示了这些解释是通过基于提示的生成过程得出的。具体而言，我们构建了一种语音条件的提示格式，将文字记录、其相关的情感标签和推理提示输入到一个大型语言模型中。我们采用两种不同的提示策略：“总结”提示鼓励基于更广泛的隐含上下文进行解释性推理，而“详细阐述”提示指导大型语言模型根据文字记录中的明确线索生成基于证据的理由。生成的解释作为监督信号，用来教导 AudioLLM 将情感类别与有意义的语言和上下文线索联系起来，使得生成的 AudioLLM 达到更具解释性和对上下文敏感的情感理解。

利用生成的目标，我们通过从一个精心设计的池子中抽取问题来构建问答训练数据，该池子旨在探测情感理解。这些问题集中在说话者的情感状态，例子包括：“你如何从他们的言语中解读说话者的情感状态？”，“你认为说话者在表达什么情感？”，以及“你会如何描述说话者声音的语气？”我们在查询粗粒度情感时采用类似的方法，使用更广泛的问题来引出说话者整体积极、消极或中立的倾向。

3.2 双编码器特征融合

我们使用的多任务 AudioLLM 框架包括：(1) 一个双编码器架构，其由一个通用语音编码器 E_{speech} 和一个情感中心编码器 E_{emotion} 组成，每个编码器设计用于捕捉音频输入的不同方面；(2) 一对轻量级适配器模块，将编码器输出投射到一个共享的潜在空间；以及 (3) 一个预训练的 LLM，它使用融合的表达生成自由形式的文本输出。情感中心编码器通过引入特定于情感线索的归纳偏差来增强情感理解和推理能力。双编码器框架的概述如图 3 所示。

我们将数据集记为 $(\mathcal{A}, \mathcal{T}, \mathcal{Y})$ ，其中 $\mathcal{A}$ 是输入音频信号的集合， $\mathcal{T}$ 是对应文本指令的集合， $\mathcal{Y}$ 是输出文本响应的集合。给定第 i 个训练样本的音频输入 $a_i \in \mathcal{A}$ ，我们提取两种类型的音频嵌入：语音编码器产生 $z_i^{\text{speech}} = E_{\text{speech}}(a_i)$ ，情感中心编码器产生 $z_i^{\text{emotion}} = E_{\text{emotion}}(a_i)$ 。我们研究了不同的情感中心编码器选择，如第 4 节所述，并在我们的实验中将 Whisper (Radford et al., 2023)（一个广泛使用的自动语音识别 (ASR) 模型）固定为语音编码器。话语被零填充到 30 秒，并且编码器嵌入的序列长度为 1500。这些编码器嵌入通过适配器模块传递，以重塑并投射到共享的潜在空间。语音编码器嵌入的序列长度被转换为 100，而情感编码器嵌入的序列长度被转换为较短的 10，强调

以最小冗余为补充信号的凝聚情感特定表示。在适配器模块的实现中，我们遵循 MERaLiON-AudioLLM (MERaLiON Team, 2024)：我们在多个时间步长上连接编码器嵌入以减少序列长度，然后通过具有两个隐藏层和 SiLU 激活的多层感知器 (MLP) 传递它们。得到的音频令牌序列为 $\text{token}_{a_i}^{\text{speech}}$ 和 $\text{token}_{a_i}^{\text{emotion}}$ 。我们在序列维度上连接这些音频令牌序列：

$$\text{token}_{a_i} = \text{token}_{a_i}^{\text{speech}} \oplus_s \text{token}_{a_i}^{\text{emotion}}.$$

我们单独对文本指令 $t_i \in \mathcal{T}$ 进行标记为 $\text{token}_{t_i} = \text{tokenizer}(t_i)$ 。然后将音频和文本标记在序列维度上进行串联：

$$\text{token}_i = \text{token}_{a_i} \oplus_s \text{token}_{t_i}.$$

最后，将连接后的标记输入到 LLM 中以生成目标响应：

$$\hat{y}_i = \text{LLM}(\text{token}_i).$$

3.3 任务交替训练

为了确保每个编码器专注于各自的任務，我们采用了一种任务交替的训练策略，如图 3 所示。具体而言，语音编码器及其适配器在以语音为中心的任务（例如，口语问答、自动语音识别）中进行训练，而情感为中心的编码器及其适配器则在以情感为中心的任务（例如，带解释生成的情感识别）中进行训练。在每轮训练中，仅更新对应当前任务的编码器及其相关的适配器以及 LLM LoRA 参数，而另一个编码器和适配器保持冻结状态。这种交替的方案实现了语音和情感表示的解耦但互补的学习。在最后一个周期中，我们会更新跨所有任务的所有适配器和 LLM LoRA 参数，以增强多模态对齐。

4 实验设置

4.1 模型实现

我们使用 Gemma-2-9B-IT (Gemma Team et al., 2024) 作为情感原因提取和 AudioLLM 框架中的 LLM。对于每个编码器，我们采用 Whisper-Large-v3 (Radford et al., 2023) 的编码器模块，这是现有 AudioLLMs (Tang et al., 2024; Chu et al., 2023; Hu et al., 2023; MERaLiON Team, 2024) 中的一个流行选择。对于以情感为中心的编码器，我们还尝试其他选择，例如更小版本的 Whisper、HuBERT (Hsu et al., 2021) 以及 Emotion2Vec (Ma et al., 2024)。对于每个适配器，我们使用一个具有两个隐藏层

的轻量级多层感知机 (MLP) 和 SiLU 激活函数, 如 MERaLiON-AudioLLM (MERaLiON Team, 2024) 中所述。我们在 8 个 H100 GPU 上进行了批量大小为 48 的多任务训练, 为期 5 个 epoch, 使用 AdamW 优化器, 其参数为 $\beta_1 = 0.9$ 和 $\beta_2 = 0.999$, 学习率为 5×10^{-5} 。用于 LLM 输入的提示模板形式为:

“< 音频 _ 开始 > { 音频标记 } < 音频 _ 结束 > { 文本指令提示 }”

4.2 数据集

我们在两个广泛使用的情感识别 (ER) 和情感倾向识别 (SR) 基准上进行训练和评估: IEMOCAP (Busso et al., 2008) 和 MELD (Poria et al., 2019)。IEMOCAP 数据集包含职业演员之间的对话, 其中每个话语都被标注为十种类别情感标签之一, 即愤怒、快乐、中性、悲伤、厌恶、恐惧、惊讶、挫败、兴奋和其他。MELD 是一个从电视节目《朋友》中提取的多模态数据集, 包含音频、视频和文本用于多方对话。在 MELD-ER 中, 每个话语被标注为七种情感类别之一, 分别是中性、快乐、厌恶、悲伤、惊讶、愤怒和恐惧。我们还包括用于情感倾向识别的 MELD-SR, 每个话语被标注为积极、消极或中性, 以评估模型在口语语境中捕捉整体情感倾向的能力。

对于语义任务, 我们利用 MNSC (Wang et al., 2025) 中的口语问答 (SQA) 任务, 这是一个以 Singlish 为中心的语料库, 这种克里奥尔语根植于英语。我们选择 MNSC 是因为预训练的编码器和 LLM 在训练中不太可能接触过其语言模式, 从而减少了因先前接触导致的性能偏差风险。为进一步分析, 我们在第 6 节中尝试了额外的 SQA 任务, 例如 Spoken-SQuAD (Lee et al., 2018) 和 SLUE-P2-SQA5 (Shon et al., 2023), 以及自动语音识别 (ASR) 任务, 如 MNSC ASR (Wang et al., 2025) 和 LibriSpeech (Panayotov et al., 2015)。

更多的数据细节和统计信息见附录 A.1。

4.3 评估

我们在第 4.2 节的数据集上使用 AudioBench (Wang et al., 2024a) 进行模型评估, 并遵循其训练-测试划分以防止数据污染。语音识别 (ASR) 任务通过词错误率 (WER) 进行评估, 剩余任务则使用 LLM-as-a-Judge 框架进行评估。模型输出由 Llama-3-70B-Instruct (AI@Meta, 2024) 根据给定的评分标准进行评估, 然后将得分标准化为 0-100 的范围, 其中更高的分数反映更好的性能。

对于情感和情感识别, 我们在二进制尺度上对每个情感预测进行评分, 其中得分为 1 表

Training Targets	IEMOCAP	MELD-ER	MELD-SR	Avg
Label Only (Original)	18.6	47.9	48.1	38.2
Interpretive Reasoning	60.8	52.6	60.1	57.8
Evidence-Grounded Reasoning	58.6	54.1	61.6	58.1

Table 1: 在不同监督目标上训练的 AudioLLM (使用 Whisper-Large-v3 编码器) 的情感和情绪识别性能。使用增强推理的目标进行训练可显著提升性能。

示与真实标签语义对齐。由于 AudioLLMs 生成开放式回答, 传统的精确匹配等指标可能不够。将 LLM 作为裁判的方法允许我们以更灵活和符合人类对齐的方式评估模型输出的事实正确性和相关性。

我们扩展了 LLM-as-a-Judge 框架来评估模型基于证据的情感推理质量。我们提取模型预测中生成的直接引用, 并使用两个关键指标对其进行评估: 基于证据的评分和相关性评分。基于证据的评分评估模型提取的引用与真实记录的对齐程度, 即它们是直接引用、忠实改写还是幻觉生成。相关性评分则衡量提取的引用在多大程度上支持标注的情感标签。对于每一个预测, 我们为 Llama-3-70B-Instruct 模型的判断提供了真实的情感标签、讲话记录以及提取的引用, 并指示其根据一个结构化的评分准则为每个标准打分, 分数范围为 0 到 2。这种评估捕捉了模型输出的事实对齐性和情感可解释性。所使用的评估提示参见附录 A.3。

5 结果

5.1 推理增强训练目标的有效性

我们使用一个配备单一 Whisper-Large-v3 编码器的基础 AudioLLM 来评估不同监督目标对情感和情绪识别性能的影响。从表 1 中的结果可以看出, 模型在使用包含情感标签和自然语言解释的推理增强目标进行训练时, 其表现始终优于仅使用标签监督进行训练的模型, 平均提升近 20%。这表明, 包含语义丰富和解释性强的目标不仅增强了可解释性, 还显著提高了识别能力, 突出显示了在上下文推理中为预测提供依据的有效性。我们在附录 A.2 中提供了不同模型预测的示例。

5.2 双编码器特征融合与训练的有效性

除了 IEMOCAP 和 MELD, 我们还在 MNSC SQA 第 3-6 部分上进行训练和评估, 以评估我们提出的方法在非情感中心任务上的影响。表格 2 展示了关键设计选择的系统比较, 包括模型架构、特征融合策略和任务交替训练。我们在双编码器架构中使用 Whisper-Large-v3 作为两个编码器。我们观察到

Concat Dim	Training	ER/SR				SQA					Overall Avg
		IEMOCAP	MELD-ER	MELD-SR	Avg	Part 3	Part 4	Part 5	Part 6	Avg	
None	Joint	43.1	51.9	61.9	52.3	49.4	48.4	57.6	62.4	54.5	53.5
	Joint	40.4	51.5	62.6	51.5	47.2	50.6	58.6	63.8	55.1	53.5
Feature	Alt 1 epoch	53.7	54.6	62.0	56.8	49.2	50.0	56.4	63.0	54.7	55.6
	Alt 4 epochs	48.5	52.8	61.4	54.2	42.4	40.4	57.0	62.4	50.6	52.1
Sequence	Joint	44.1	48.8	58.1	50.3	32.8	30.8	40.2	47.6	37.9	43.2
	Alt 1 epoch	56.6	52.4	60.7	56.6	48.0	50.6	59.2	63.4	55.3	55.8
	Alt 4 epochs	55.1	52.0	61.5	56.2	52.4	49.4	57.6	64.4	56.0	56.1

Table 2: 不同方法在双编码器特征融合和多任务训练中的性能比较。Concat Dim "None" 表示单编码器基线。"Joint" 表示所有任务在所有编码器 + 适配器集合上一起训练 5 个 epoch。"Alt x epoch(s)" 指的是交替训练以语音为中心和以情感为中心的任务，每隔 x 个 epoch 在其各自的编码器 + 适配器上进行，最多到 4 个 epoch 的数据，然后在最后一个 epoch 对所有任务使用所有适配器进行训练。

Training	PL Encoder	ER/SR				SQA					Overall Avg
		IEMOCAP	MELD-ER	MELD-SR	Avg	Part 3	Part 4	Part 5	Part 6	Avg	
Alt 1 epoch	Whisper-Large (637M)	56.6	52.4	60.7	56.6	48.0	50.6	59.2	63.4	55.3	55.8
	Whisper-Small (88M)	56.8	52.5	60.6	56.6	50.8	50.4	58.6	61.4	55.3	55.9
	Whisper-Tiny (8M)	46.2	49.3	53.1	49.5	25.8	23.8	27.0	32.0	27.2	36.7
	HuBERT-XL (962M)	51.3	46.1	51.9	49.8	26.8	23.6	27.0	28.6	26.5	36.5
	Emotion2Vec+ Large (164M)	57.1	50.0	56.6	54.6	28.2	22.4	28.2	32.3	27.8	39.3
	Emotion2Vec+ base (93M)	63.5	45.9	55.0	54.8	24.6	24.2	29.4	29.6	27.0	38.9
Alt 4 epochs	Whisper-Large (637M)	55.1	52.0	61.5	56.2	52.4	49.4	57.6	64.4	56.0	56.1
	Whisper-Small (88M)	55.3	52.6	61.6	56.5	48.6	49.2	61.4	61.8	55.3	55.8
	Whisper-Tiny (8M)	50.3	52.7	59.7	54.2	47.8	47.0	58.0	62.2	53.8	54.0
	HuBERT-XL (962M)	44.0	49.3	59.0	50.8	49.6	47.8	59.0	62.2	54.7	53.0
	Emotion2Vec+ Large (164M)	63.8	53.0	61.1	59.3	49.2	51.2	59.0	62.8	55.6	57.2
	Emotion2Vec+ base (93M)	56.3	52.0	60.4	56.2	46.2	45.8	54.4	61.0	51.9	53.7

Table 3: 双编码器架构中不同情感中心编码器选择的性能比较。“Alt x epoch(s)” 指的是每 x 时代交替训练以语音为中心和以情感为中心的任务在各自的编码器 + 适配器上，最多达到 4 个时代的数据，然后在最后一个时代对所有任务在所有适配器上进行训练。

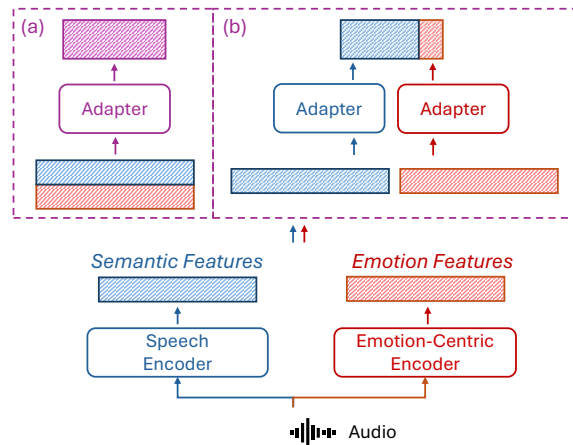


Figure 4: 特征融合策略：我们探索了 (a) 沿特征维度的融合，其中来自两个编码器的特征按通道方向连接，以及 (b) 沿序列维度的融合，其中跨时间步的特征按令牌方向连接。

1. 双编码器架构可以胜过单编码器基准，这表明结合互补表示可以提升整体性能；
2. 沿序列维度连接相比沿特征维度连接能得到稍好的结果，这可能是由于更好地保留了时间结构。在后者中，语音和情感嵌入都被重塑为长度 100，沿特征维度连接，

然后通过一个适配器模块。这两种连接方式如图 4 所示；

3. 与联合多任务训练相比，交替训练任务可以提高性能，特别是在情感和情绪识别方面。

在表 3 中，我们进一步比较了不同情感中心编码器选择上的多任务表现。我们采用了一个双编码器架构，其中特征沿序列维度进行拼接，并应用任务交替训练。我们的观察包括：

1. 每轮任务特定训练必须足够长，以确保模型收敛；
2. 编码器的选择对性能有显著影响。具体来说，Emotion2Vec+ Large 经过情感识别的预训练，为以情感为中心的任务提供了更相关的特征，从而提升了情感相关和整体的性能。

5.3 情感推理响应的质量

表 4 展示了我们 AudioLLM 在响应中证据依托推理的评估。跨越所有评估的数据集，模型的响应中超过 49 % 明确包含了直接引用的语音内容，作为推理或解释的支持性证据。模型在扎根性得分上取得了平均 82.8 % 的高分，这表明大多数引用对原始语音记录是忠实的。

Scores	IEMOCAP	MELD-ER	MELD-SR	Avg
Quotation	73.8	49.2	49.6	57.5
Groundedness	90.6	79.2	78.5	82.8
Relevance	71.9	60.0	64.1	65.3

Table 4: 通过基于证据的推理预测评估模型预测。引用分数衡量包含至少一个可提取引文的预测百分比。基础性和相关性分数（0–100 分等级）分别评估与记录文本的一致性以及对真实情感标签的支持。

Model	IEMOCAP	MELD-ER	MELD-SR	Avg
WavLLM	59.8	41.6	51.1	50.8
Qwen2-Audio	54.0	41.6	53.9	49.8
Cascade: Whisper+SEA-LION	44.3	47.4	56.6	49.4
Phi-4-Multimodal	41.0	43.5	51.6	45.4
MERaLiON	48.5	36.4	46.2	43.7
Cascade: Whisper+Llama3	46.7	36.8	45.6	43.0
Qwen-Audio	29.4	50.7	44.9	41.7
SALMONN	23.8	30.5	41.8	32.0
AudioLLM-Reasoning	63.8	53.0	61.1	59.3

Table 5: 端到端 AudioLLM 和级联模型的情感和情绪识别性能。

相关性得分平均为 65.3%，这表明大多数引用有意义地支持了真实的情感标签，尽管特别是在 MELD 数据集上仍有改进余地。

5.4 与其他模型的比较

我们将情感和情绪识别表现与在 AudioBench 中评估的端到端 AudioLLM 进行比较：WavLLM (Hu et al., 2023)、Qwen2-Audio-7B-Instruct (Chu et al., 2024)、Phi-4-Multimodal-Instruct (Abouelenin et al., 2025)、MERaLiON-AudioLLM (MERaLiON Team, 2024)、Qwen-Audio-Chat (Chu et al., 2023) 和 SALMONN (Tang et al., 2024)。我们还与级联模型进行比较，这些模型通过在将转录内容输入大型语言模型之前，使用自动语音识别模块将音频转换为文本，分阶段处理语音：使用 SEA-LIONv3 的 Whisper-Large-v2 (Ng et al., 2025)，以及使用 Llama-3-8B-Instruct 的 Whisper-Large-v3 (AI@Meta, 2024)。从表格 5 来看，我们提出的与推理增强目标一起训练的 AudioLLM 在 IEMOCAP、MELD-ER 和 MELD-SR 上实现了最佳表现。

我们在表格 6 中比较了 AudioLLM 与情感分类器 Emotion2Vec+ Large 在 MELD 上的表现。使用推理增强目标训练的 AudioLLM 取得了最高的加权平均准确率 (53.0%)，超过了仅用标签训练的变体 (47.1%) 和 Emotion2Vec+ Large (44.7%)。然而，各模型的非加权平均值仍然相似，这表明像厌恶和恐惧这样罕见的类别仍然具有挑战性。

需要注意的是，与现有模型的比较没有控制训练数据的差异，这可能会影响性能结果。

Class	Num Samples	Emotion2Vec+ Large	AudioLLM - Label Only	AudioLLM - Reasoning
Neutral	1256	54.2	64.7	83.8
Joy	402	54.5	41.5	36.3
Disgust	68	0.0	8.8	13.2
Sadness	208	32.2	23.1	15.9
Surprise	281	38.8	28.8	26.0
Anger	345	26.1	32.2	31.3
Fear	50	2.0	6.0	8.0
Unwt Avg	2610	29.7	29.3	29.2
Wt Avg	2610	44.7	47.1	53.0

Table 6: Emotion2Vec+ Large 分类器与 AudioLLM (使用 Emotion2Vec+ Large 情感中心编码器) 在 MELD-ER 上基于标签或推理目标进行训练的逐类情感识别性能比较。非加权平均 (Unwt Avg) 无视样本大小，平等对待所有类别，而加权平均 (Wt Avg) 通过相应加权每个类别来反映样本分布。

Model	ER/SR (↑)	SQA (↑)	ASR (↓)
Base AudioLLM	44.1	56.0	19.5
+ Emotion Supervision	56.4	54.1	19.6

(a) 基础 AudioLLM 在 MNSC SQA 第 3-6 部分和 MNSC ASR 第 3-6 部分上进行了训练。

Model	ER/SR (↑)	SQA (↑)	ASR (↓)
Base AudioLLM	35.6	80.3	3.8
+ Emotion Supervision	58.0	79.0	3.6

(b) 基础 AudioLLM 在 Spoken-SQuAD 和 SLUE 上进行 SQA 训练，并在 LibriSpeech Clean 和 Other 数据集上进行 ASR 训练。

Table 7: 在已训练的基础 AudioLLM 上增加情感监督的效果。增加情感监督提升了情感理解能力，但对其他任务的性能略有影响。

6 进一步分析

我们进一步研究能否将情感理解能力引入到一个最初并未为这些任务训练的模型中。从一个没有任何情感特定监督的基础 AudioLLM 开始，我们探索在架构中添加一个以情感为中心的编码器。我们训练以情感为中心的编码器 Emotion2Vec+ Large、适配器和 LLM LoRA 在以情感为中心的任务上，然后微调适配器和 LLM LoRA 在所有任务上。

表 7 显示了将情感监督引入在不同上游任务上训练的基础 AudioLLM 的效果。在这两种训练配置中，我们观察到了在情感和情绪识别方面显著的提升，分别提高了 12.3 和 22.4 个点。这种情感理解的增强伴随着模型在原始任务上成绩的轻微下降。例如，SQA 的表现保持相当，ASR 的表现大体保持或略有改善。这些结果表明，情感能力可以有效地注入到多模态模型中，而不会牺牲其现有的能力。

7 结论

在这项工作中，我们提出了一个统一的框架，将情感推理引入多任务的音频大语言模型，结合双编码器、增强推理的监督和任务交替训练。我们的方法提高了情感和情绪识别的准确性，并能够生成基于证据的解释，如在 IEMOCAP 和 MELD 基准数据集上的表现所示。该工作突出了生成型音频大语言模型在更具解释性和情感感知的语音理解中的潜力。

虽然我们的框架显著提高了 AudioLLMs 在情感识别和解释能力方面的表现，但仍然存在一些限制。解释生成的质量可能会因情感和说话者风格的不同而有所变化，尤其是对于微妙或模糊的情感状态。此外，提取的情感推理的质量依赖于教师 LLM 的能力，这可能会带来监督中的偏见或不准确。目前的情感推理基准在多样性和规模上存在局限，凸显了需要更全面的评估数据集，以捕捉更广泛的情感表达和丰富的上下文。

我们研究中使用的数据集，包括 IEMOCAP 和 MELD，都是公开可用且在研究界广泛使用的。然而，我们强调，在现实世界或敏感环境下部署情感感知人工智能系统时，需要谨慎，因为对情感线索的误解可能导致意想不到的后果。在学术环境之外应用这些技术时，确保透明度、用户同意和适当的保护措施是至关重要的。

8

致谢 本研究得到了新加坡国家研究基金会在其国家大型语言模型资助计划下的支持。本文中表达的任何观点、研究结果、结论或建议均为作者（们）的观点，并不代表新加坡国家研究基金会的立场。

References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yiling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and Xiren Zhou. 2025. Phi-4-Mini technical report: Compact yet powerful multimodal language models via mixture-of-LoRAs. *arXiv*.
- AI@Meta. 2024. [Llama 3 model card](#).
- Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. SD-Eval: A benchmark dataset for spoken dialogue understanding beyond words. *arXiv*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, and et al. 2023. Qwen technical report. *arXiv*.
- Jaime Bellver, Ivan Martín-Fernández, Jose Bravo-Pacheco, Sergio Esteban, Fernando Fernández-Martínez, and Luis D’Haro. 2024. Multimodal audio-language model for speech emotion recognition. In *Odyssey: The Speaker and Language Recognition Workshop*.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. [IEMOCAP: Interactive emotional dyadic motion capture database](#). *Language Resources and Evaluation*, 42:335–359.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, and et al. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *ICML*.
- 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](#).
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-Audio technical report. *arXiv*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-Audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv*.
- Nilaksh Das, Saket Dingliwal, S. Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2024. SpeechVerse: A large-scale generalizable audio language model. *arXiv*.
- Alexandre D’efosse, Laurent Mazar’e, Manu Orsini, Am’elie Royer, Patrick P’erez, Herv’e J’egou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv*.
- Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *NeurIPS*.

- Yumeng Fu, Junjie Wu, Zhongjie Wang, Meishan Zhang, Lili Shan, Yulin Wu, and Bingquan Liu. 2025. LaERC-S: Improving LLM-based emotion recognition in conversation with speaker characteristics. In *COLING*.
- Morgane Riviere Gemma Team, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, and et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv*.
- William Held, Minzhi Li, Michael Ryan, Weiyan Shi, Yanzhe Zhang, and Diyi Yang. 2024. Distilling an end-to-end voice assistant without instruction training data. *arXiv*.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.
- Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linqun Liu, and Furu Wei. 2023. WavLLM: Towards robust and adaptive speech large language model. *arXiv*.
- Shengpeng Ji, Yifu Chen, Minghui Fang, Jialong Zuo, Jingyu Lu, Hanting Wang, Ziyue Jiang, Long Zhou, Shujie Liu, Xize Cheng, Xiaoda Yang, Zehan Wang, Qian Yang, Jian Li, Yidi Jiang, Jingzhen He, Yunfei Chu, Jin Xu, and Zhou Zhao. 2024. WavChat: A survey of spoken dialogue models. *arXiv*.
- Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, and Ozlem Kalinli. 2024. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv*.
- Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. 2018. Spoken SQuAD: A study of mitigating the impact of speech recognition errors on listening comprehension. In *Interspeech*.
- Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, and Jian Luan. 2025. Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering. *arXiv*.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. 2024a. DeSTA: Enhancing speech language models through descriptive speech-text alignment. In *Interspeech*.
- Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, Chao-Han Huck Yang, Jagadeesh Balam, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung-yi Lee. 2024b. DeSTA2: Developing instruction-following speech language model without speech instruction-tuning data. In *ICASSP*.
- Ziyang Ma, Zhuo Chen, Yuping Wang, Eng Siong Chng, and Xie Chen. 2025. Audio-CoT: Exploring chain-of-thought reasoning in large audio language model. *arXiv*.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. 2024. emotion2vec: Self-supervised pre-training for speech emotion representation. In *ACL*.
- MERaLiON Team. 2024. MERaLiON-AudioLLM: Bridging audio and language with large language models. *arXiv*.
- Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Ngee Chia Tai, Wai Yi Leong, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulgadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Wayne Lau, Choon Meng Tan, and 12 others. 2025. SEA-LION: Southeast Asian Languages in One Network. *arXiv*.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*.
- 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 *ACL*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.
- S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2025. MMAU: A massive multi-task audio understanding and reasoning benchmark. In *ICLR*.
- Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei Yu Wu, Hung yi Lee, Karen Livescu, and Shinji Watanabe. 2023. SLUE Phase-2: A benchmark suite of diverse spoken language understanding tasks. *ACL*.
- Chuanneng Sun, Zeeshan Ahmed, Yingyi Ma, Zhe Liu, Lucas Kabela, Yutong Pang, and Ozlem Kalinli. 2024. Contextual biasing of named-entities with large language models. In *ICASSP*.
- Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *ICLR*.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, and et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. AudioBench: A universal benchmark for audio large language models. *NAACL*.

Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F Chen, and AiTi Aw. 2025. Advancing Singlish understanding: Bridging the gap with datasets and multimodal models. *arXiv*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024b. BLSP-Emo: Towards empathetic large speech-language models. In *EMNLP*.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024. Beyond silent letters: Amplifying LLMs in emotion recognition with vocal nuances. *arXiv*.

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-Reasoner: Improving reasoning capability in large audio language models. *arXiv*.

Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. 2024. SECAP: speech emotion captioning with large language model. In *AAAI*.

Zhixian Zhao, Xinfa Zhu, Xinsheng Wang, Shuiyuan Wang, Xuelong Geng, Wenjie Tian, and Lei Xie. 2025. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. *arXiv*.

A 实验细节

A.1 数据集

我们在两个广泛使用的情感识别 (ER) 和情绪识别 (SR) 基准上进行训练和评估: IEMOCAP (Busso et al., 2008) 和 MELD (Poria et al., 2019)。IEMOCAP 是根据定制的非商业研究许可提供的, 而 MELD 是根据 GNU 通用公共许可证 v3.0 (GPL-3.0) 分发的。由于 IEMOCAP 缺乏预定义的训练-测试划分, 我们采用 AudioBench (Wang et al., 2024a) 中定义的 90-10 划分, 包含 9035 个训练样本 (愤怒: 1140, 厌恶: 2, 兴奋: 1816, 恐惧: 98, 挫折: 2608, 快乐: 588, 中立: 1539, 其他: 23, 悲伤: 1120, 惊讶: 101) 和 1004 个测试样本 (愤怒: 129, 厌恶: 0, 兴奋: 160, 恐惧: 9, 挫折: 309, 快乐: 68, 中立: 187, 其他: 3, 悲伤: 130, 惊讶: 9)。MELD 有 9988 个训练样本 (愤怒: 1109, 厌恶: 271, 恐惧: 268, 快乐: 1743, 中立: 4709, 悲伤: 683, 惊讶: 1205) 和 2610 个测试样本 (愤怒: 345, 厌恶: 68, 恐惧: 50, 快乐: 402, 中立: 1256, 悲伤: 208, 惊讶: 281)。

对于语义任务, 我们使用了多任务国家语音语料库 (MNSC) (Wang et al., 2025), 特别是于新加坡开放数据许可证下发布的 SQA 第 3-6 部分和 ASR 第 3-6 部分。我们还使用了在 CC-BY-SA-4.0 许可证下发布的 Spoken-SQuAD (Lee et al., 2018), SLUE-P2-SQA5 (Shon et al., 2023), 这是在 CC-BY-SA-4.0 许可证和 Apache 许可证 2.0 下发布的一系列数据集, 以及在 CC-BY-4.0 许可证下发布的 LibriSpeech (Panayotov et al., 2015)。

本文中的所有实验均遵循数据集的相应许可和使用条款。

A.2 AudioLLM 响应

我们对 IEMOCAP、MELD-ER 和 MELD-SR 数据集上的不同推理格式进行模型预测的定性分析。每个示例都包括文字记录、真实标签和模型生成的输出, 输出基于三种监管类型: 仅标签、解释性推理和基于证据的推理。如表 8、9 和 10 所示, 仅标签的响应通常无法捕捉到正确的情感细微差别, 即使情感显而易见, 也会默认为中性预测。相比之下, 解释性推理和基于证据的推理与真实情况更一致, 提供更丰富的理由和更好的情感识别能力。值得注意的是, 基于证据的推理通过明确地将语音内容和情感线索与预测的情感联系起来, 展示了更高的清晰度。

A.3 评估

我们遵循 AudioBench 的 LLM-as-a-Judge 框架来评估模型的基于证据的情感推理质量。我们提取模型预测中的直接引用, 并使用两个关键指标来评估它们: 扎根性分数和相关性分数。对于每个预测, 我们为 Llama-3-70B-Instruct 模型法官提供真实的情感标签、语音转录和提取的引用, 并指示其根据给定的结构化评分标准为每个标准分配 0 到 2 的分数。然后将这些分数归一化为 0-100 的范围。使用的评估提示如图 5 所示。

Evaluation Prompt Template

```
[Ground Truth Emotion]
{reference}

[Ground Truth Transcript]
{transcript}

[Extracted Quotes from Model Prediction]
{extracted_quotes}

[Evaluation Task]
Evaluate the extracted quotes using the following three criteria.

**Groundedness Score**
Assess whether the extracted quotes are grounded in the ground truth transcript.
Scoring Guide:
Score0: The quotes do not appear in the ground truth transcript and are not
semantically aligned (i.e., hallucinated or generic).
Score1: The quotes partially match the ground truth transcript. There may be
loose paraphrasing or selective grounding.
Score2: The quotes are clearly derived from the ground truth transcript, through
direct quotes or faithful paraphrases.

**Relevance Score**
Assess whether the extracted quotes support the ground truth emotion label.
Scoring Guide:
Score0: The quotes are irrelevant or inconsistent with the ground truth emotion.
They may even suggest a different emotion.
Score1: The quotes are loosely related to the ground truth emotion but lack
clarity, specificity, or completeness.
Score2: The quotes clearly and directly support the ground truth emotion.

Respond with the following structured format:

Ground Truth Emotion: (string)
Ground Truth Transcript: (string)
Extracted Quotations from Model Prediction: (list of strings)
Groundedness Score: (int)
Relevance Score: (int)
Explanation: (string - justify the assigned scores)
```

Figure 5: 用于评估提取的情感证据的基础性和相关性的评估提示。

Example 1 Transcript: Is he an actor too, or-? Ground Truth Emotion Label: Neutral Label Only: The speaker's speech reflects a neutral state Interpretive Reasoning: The speaker is asking if someone was an actor too, implying a shared experience or connection. They seem neutral in their tone, neither excited nor disappointed. Evidence-Grounded Reasoning: The speaker's tone is neutral, suggesting they are neither expressing strong positive nor negative emotions. They are simply seeking clarification or information. The phrase "Is he an actor too?" implies a previous context where someone was identified as an actor. The speaker is now inquiring if another person also shares this profession.
Example 2 Transcript: Oh, I will never forget his face. Ooh-rah [LAUGHTER]. Ground Truth Emotion Label: Excited Label Only: The speaker's speech suggesting a neutral state. Interpretive Reasoning: The speaker is laughing and reminiscing about someone's face, implying a positive and humorous memory. They express excitement and amusement. Evidence-Grounded Reasoning: The speaker is clearly excited and perhaps a little bit nostalgic. Here's why: "Oh, I will never forget his face": This phrase suggests a strong memory, likely positive given the excited tone. Laughter: Laughter often accompanies happy memories and a sense of lightheartedness. The speaker is reminiscing about someone and a specific moment, and the excitement suggests that the memory is a pleasant one.
Example 3 Transcript: You've got to tell me. Ground Truth Emotion Label: Frustration Label Only: The speaker's speech suggesting a neutral state. Interpretive Reasoning: The speaker is frustrated and wants the listener to tell them something. Evidence-Grounded Reasoning: The speaker is frustrated. They are likely feeling impatient, annoyed, or even angry because they feel like something is not being done or explained properly. The phrase "You've got to tell me" suggests a sense of urgency and a demand for information.

Table 8: IEMOCAP: 情感预测的定性比较。在评估数据集中呈现了前三个样本。每个示例包括转录文本、真实情感标签和模型预测及其推断的情感。情感标签显示在 **blue** 中，语音内容的引文显示在 **orange** 中。解释性和证据支撑的推理与真实情况更一致，提供了更丰富的理由和改进的情感识别。

<u>Example 1</u> Transcript: Yeah, I'm not in that. Ground Truth Emotion Label: Neutral Label Only: The speaker's emotion seems neutral. Interpretive Reasoning: The speaker is denying being in something, and they sound neutral about it. Evidence-Grounded Reasoning: The speaker's tone is neutral, indicating they are not experiencing any strong emotions. They are simply stating a fact without any particular emphasis or feeling behind their words.
<u>Example 2</u> Transcript: He's a dropper! Ground Truth Emotion Label: Anger Label Only: The speaker's emotion seems neutral. Interpretive Reasoning: The speaker is expressing anger and frustration towards someone they perceive as a "dropper," likely someone who is unreliable or lets them down. Evidence-Grounded Reasoning: The speaker's tone is neutral, but the statement itself carries a strong implication. While the speaker doesn't express anger or disgust directly, the phrase "He's a dropper!" suggests a negative judgment about the person being referred to. It implies that this person is unreliable, perhaps someone who makes promises they don't keep or fails to deliver on expectations.
<u>Example 3</u> Transcript: Yes! Ground Truth Emotion Label: Joy Label Only: The speaker's speech reflects neutral. Interpretive Reasoning: The speaker said "Yes." The emotion conveyed is neutral. Evidence-Grounded Reasoning: The speaker's tone is neutral, indicating they are not experiencing any strong positive or negative emotions. They are likely calm, composed, and objective in their state of mind.

Table 9: MELD-ER: 情绪预测的定性比较。展示了评估数据集中的前三个样本。每个示例包括讲话稿、真实情绪标签以及模型预测及其推测的情绪。情绪标签在 blue 中，讲话内容的引用在 orange 中。解释性和以证据为基础的推理通过利用讲话内容，提供了更丰富的讲话者情绪解释。

<u>Example 1</u> Transcript: Yeah, I'm not in that. Ground Truth Emotion Label: Neutral Label Only: The speaker's sentiment is neutral. Interpretive Reasoning: The speaker is stating that they are not in a particular location or situation. They express this in a neutral tone, indicating no strong feelings about the matter. Evidence-Grounded Reasoning: The speaker's tone is neutral, indicating they are not expressing strong positive or negative emotions. They are simply stating a fact without any noticeable inflection or emphasis.
<u>Example 2</u> Transcript: He's a dropper! Ground Truth Emotion Label: Negative Label Only: The speaker's sentiment is neutral, based on their speech. Interpretive Reasoning: The speaker is expressing a negative sentiment towards someone they perceive as a "dropper," which likely implies a negative trait or action. Evidence-Grounded Reasoning: The sentiment is negative. The speaker's tone suggests disapproval or disgust towards the person being referred to as a "dropper." The word "dropper" itself often carries a negative connotation, implying someone who is unreliable, messy, or in some way undesirable.
<u>Example 3</u> Transcript: Yes! Ground Truth Emotion Label: Positive Label Only: The speaker's sentiment is neutral, based on their speech. Interpretive Reasoning: The speaker said "Yes." The sentiment is neutral. Evidence-Grounded Reasoning: The sentiment is positive. The speaker's enthusiastic "Yes!" conveys a strong sense of agreement, excitement, or approval.

Table 10: MELD-SR: 情感预测的定性比较。展示了评估数据集中的前三个样本。每个例子都包括文字记录、真实情感标签以及模型预测及其推断的情感。情感标签用 blue 表示，讲话内容的引用用 orange 表示。解释性和基于证据的推理能够更好地与真实情况对齐，提供更丰富的论证和改进的情感识别。