人脸并非你所需的一切：面向不完整多模态情感识别的 MIME 基准

Yuxin Jia
jiayuxin@mails.neu.edu.cn 东北大学计算机科学与工程学院，中国沈阳；武汉大学湖北省多媒体与网络通信工程重点实验室，中国武汉

Xing Lan*
lanx@cse.neu.edu.cn 东北大学计算机科学与工程学院，中国沈阳；南京大学新型软件技术国家重点实验室，中国南京

Wensong Wang
20235870@stu.neu.edu.cn
计算机科学与工程，东北大学
沈阳，中国
国家多媒体软件工程技术研究多媒体软件中心（NERCMS），武汉大学
武汉 ，中国

Jian Xue
xuejian@ucas.ac.cn
工程科学学院，中国科学院大学
北京，中国

Feiliang Ren
renfeiliang@cse.neu.edu.cn
计算机科学与工程学院，东北大学
沈阳，中国

Ke Lu
luk@ucas.ac.cn
工程科学学院，中国科学院大学
中国北京

arXiv:2611.30288v1 [cs.MM] 23 Nov 2026

摘要

多模态情感识别（MER）是多媒体理解中的一项基础任务。当前最先进的方法在完整且对齐良好的多模态输入这一假设下取得了显著成功。然而，在现实世界的不受约束场景中，模态退化和信息丢失往往难以预测，导致现有多模态情感识别模型的性能灾难性崩溃，因为这些模型严重依赖面部线索作为核心推理依据。更为严峻的是，当前仍存在一个关键研究空白：鲜有系统性研究在各种细粒度的不完整信息条件下评估模型鲁棒性；既有研究仅涉及粗粒度的全模态移除，也忽视了对情绪推理过程的可解释性分析。为填补这一空白，我们正式定义了不完整多模态情感识别（IMER）任务，用于评测模型在多粒度模态和情感信息丢失条件下的泛化能力与鲁棒性。我们进一步构建了专门的 IMER基准 MIME，包含 2,000 个视频片段，涵盖真实的细粒度面部细节退化、严重面部退化以及极端全模态丢失，超越了以往研究有限的粗粒度设置。此外，我们提出了情感链（CoE）分析范式及配套评估指标，用以剖析多模态大语言模型（MLLMs）对细粒度情感线索的感知与推理过程。针对七种主流多模态大语言模型开展的大量实验表明，在面部退化时，肢体语言和场景上下文是关键的补偿机制，而模型之间仍存在显著差距基础视觉感知能力与高层次认知推理。我们的工作提供了一个聚焦的评估平台和可操作的见解，以推动稳健的多模态情感识别。我们的数据集和代码公开于 https://yuxinokk.github.io/MIME。

CCS 概念

• **计算方法** → **场景理解**; 活动识别与理解。

关键词

多模态情感识别；模态缺失；情感链；基准数据集；大型视觉语言模型

ACM 参考格式：

YuxinJia、Xing Lan、Wensong Wang、Jian Xue、Feiliang Ren 和 Ke Lu。2026。人脸并非你所需的一切：不完整多模态情感识别的 MIME 基准。载于第 *34*届 *ACM* 多媒体国际会议论文集（*MM '26*），*2026*年*11*月*10*日至*14*日，巴西里约热内卢。ACM，纽约州纽约市，美国，8页。
https://doi.org/10.1145/3767308.3838643

1 引言

多模态情感识别（MER）是多媒体领域的一项重要任务，旨在从多模态信息中推断人类情绪状态 [9, 13, 16, 42]。近期，多模态情感识别领域的进展 [1, 18, 23] 通过有效融合多种多模态信息（如面部动态、声音细微特征、肢体语言和上下文视觉线索等），取得了显著成效；这类方法通常依赖于输入的多模态信息完整且对齐良好的假设。

然而，在不受约束的现实世界部署场景中，由于各种实际因素，多模态数据往往会遭受不可预测的退化和模态缺失 [7, 24, 38, 49], , 例如严重的面部遮挡、光照波动、传感器故障、网络丢包以及隐私约束。大多数现有的多模态情绪识别方法 [19, 22, 40, 43] 高度

*通讯作者：lanx@cse.neu.edu.cn

本作品采用知识共享署名4.0国际许可协议授权。*MM '26*，里约热内卢，巴西
© 2026 版权归所有者/作者所有。ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3838643

依赖面部表情特征作为情感推断的核心和主导线索，这会导致性能灾难性下降，甚至在关键面部模态受损、失效或完全不可用时彻底丧失功能。

尽管实际场景对鲁棒多模态情感识别系统 [5, 12, 35] 的需求十分迫切，但仍存在一个显著且尚未探索的研究空白：鲜有研究系统地考察并基准评估当前多模态情感识别模型在各种细粒度、不完整信息条件下的表现。为填补这一空白，我们正式定义一项新任务：**不完整多**模态情感识别（**IMER**），重点评估情感识别模型在模态信息和情感信息以不同程度缺失的情况下的泛化能力与鲁棒性。与基于模态理想且完整这一假设的标准多模态情感识别任务相比，IMER 能更准确地反映情感识别系统在不受约束的部署环境中的实际表现和实用可靠性。

为便于开展这项任务，我们构建了专用基准**多模态情**绪中的信息缺失（**MIME**），用于评估模型在各种信息缺失场景下的表现，其中包含 2,000 个视频片段。与现有仅模拟极端粗粒度全模态缺失（例如音频或视觉流完全缺失）的 IMER基准不同，MIME 强调现实世界中的信息不完整本质上具有多种粒度：在许多实际情况下，视觉模态并未完全丢失，而是关键的面部情感证据仅遭到部分退化。因此，我们将细粒度的面部细节退化与更严重的面部退化区分开来，而非假设面部证据要么完全可用，要么彻底缺失。

此外，我们设计了一种名为情感链（**CoE**）的新型分析范式，用于剖析多模态大语言模型（MLLMs）在情感识别过程中的细粒度情感线索感知与推理过程。具体而言，CoE 明确引导模型在预测最终情绪之前，分别解耦并阐述其对各个模态的观察结果，即面部、身体、场景和音频，从而突破单一最终准确率指标的局限，为理解模型行为提供可解释的见解。此外，我们还为 CoE 制定了针对具体任务的评估指标，涵盖场景理解、情绪分析、分类准确率和整体 CoE 质量。

我们的广泛评估结果表明，尽管面部是主要的情绪指示因素，但在面部退化时，肢体语言和场景上下文可作为重要的补偿机制，有效避免性能灾难性崩溃。即使是最强大的多模态大语言模型，在面对细粒度面部信息损坏和跨模态联合缺失场景时，也会出现显著的性能下降，这暴露了现有模型在真实环境中的严重鲁棒性局限。CoE 评估揭示了模型的基础视觉感知能力与其高层认知推理技能之间存在显著差距。

我们的主要贡献总结如下：

- 我们正式定义了 IMER 任务，这是首个聚焦于多模态情绪识别中多粒度信息缺失的系统性任务构想，弥合了标准理想设置与现实部署需求之间长期存在的鸿沟。
- 我们构建了 MIME，这是首个专为 IMER 打造的基准，配备 CoE分析范式和任务专属评估指标，为研究社区提供了一个标准化且可扩展的评估平台。
- 我们在多种模态缺失条件下，对 7 个主流开源和闭源多模态大语言模型开展了全面的基准实验。我们揭示了它们的优势与局限，为未来研究推进多模态大语言模型的情绪智能铺平了道路。

2 相关工作

2.1 多模态情感识别数据集

大规模数据集从根本上推动了多模态情感识别（MER）的发展。早期的情感计算研究依赖于在实验室采集的数据集，例如 IEMO-CAP [3] 和 MELD [31]。为了捕捉更贴近自然状态的人类情绪，近期的基准已转向野外场景。例如，CAER [15]开创了情境感知情感识别，强调面部表情与背景场景的共同作用。DFEW [11] 和 MAFW [26] 提供了从真实媒体中提取的大规模动态视频片段，涵盖丰富的视听多模态线索。此外，FERV39k[41] 收集了大量不受约束的无声视频片段，推动了纯视觉情绪分析的研究。然而，这些基准存在一个根本局限：它们都依赖于模态完整这一假设。此外，其标注严格局限于离散类别标签（例如，快乐、悲伤）。它们缺少解读复杂情绪状态所需的细粒度推理依据[38, 44, 46, 49], , 这显著限制了其在训练可解释情感模型方面的效用。

2.2 模态缺失条件下的情感识别

近期研究 [8, 17, 32] 探索了模态缺失条件下的多模态情绪识别，但仅限于粗粒度设置：从标准数据集中简单移除固定模态，并仅保留 1-2 种模态。这种范式既无法支持在语义特定的信息丢失条件下进行评估，也无法建模情绪推理过程，或分析有效模态对情感识别的协同效应 [50]。

为弥补这一不足，我们构建了一个涵盖多种粒度的模态缺失 MIME 基准。与以往数据集不同，我们的基准将视觉元素独特地解耦为不同的细粒度模态，并支持全面评估模态缺失条件下的鲁棒性。

3 MIME 基准

3.1 数据采集与预处理

为确保社交情境和缺失程度的多样性，我们从社交媒体以及四个权威的自然场景基准中收集原始数据：**CAER**、**DFEW**、**MAFW**和**FERV39k**。我们重点关注

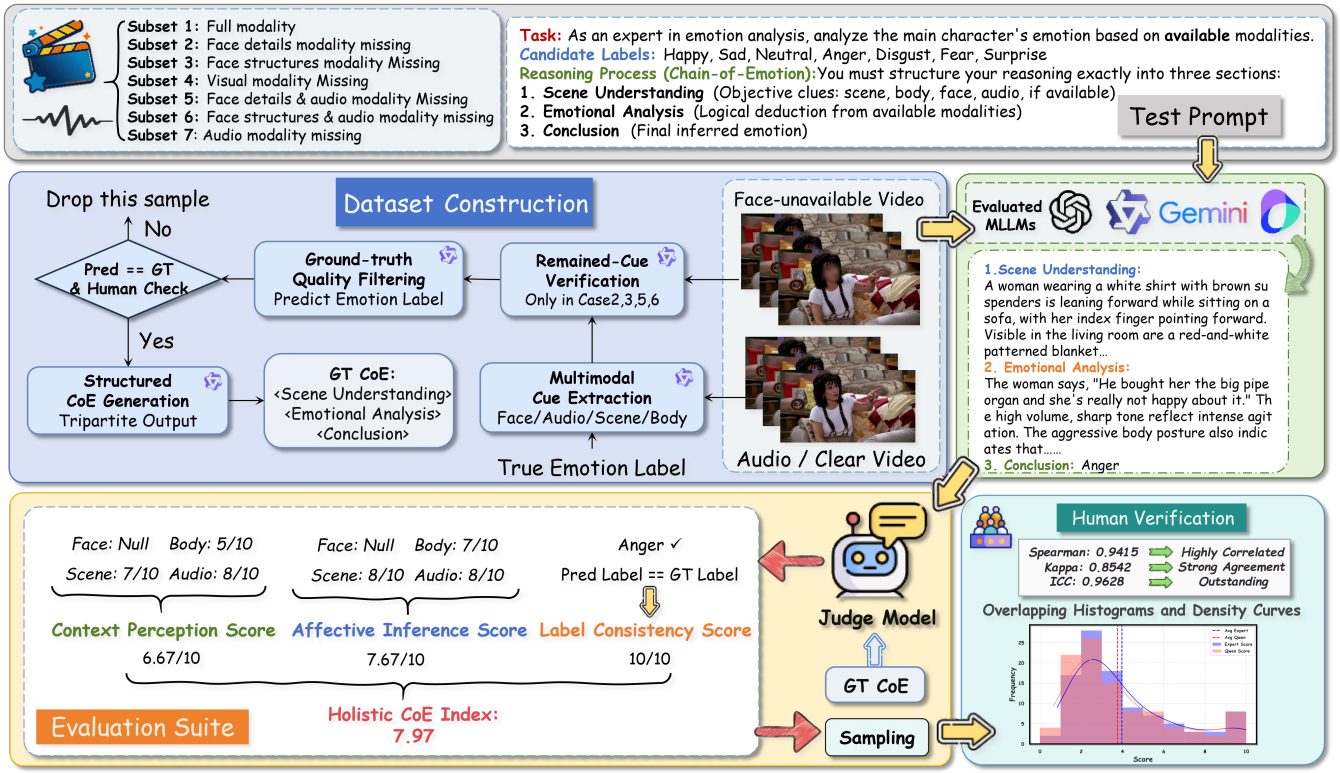

图 1：我们数据构建与评估策略的示意图。

表 1：MIME 与其他 IMER 相关基准的比较。

基准	格式	补充材料 不完整	不完粒度	情感链 情感
EmotionBench [39]	Text	✗	✗	✗
EmoBench [34]	Text	✗	✗	✗
MOSABench [36]	图像	✗	✗	✗
MM-InstructEval [45]	图像	✗	✗	✗
IEMOCAP [2]	视频	✓	粗粒度	✗
MC-EIU [25]	视频	✗	✗	✗
OV-MER [20]	视频	✗	✗	✗
EmoBench-M [10]	视频	✗	✗	✗
MER-UniBench [21]	视频	✗	✗	✗
MOSI [47]	视频	✓	粗粒度	✗
MME-Emotion [48]	视频	✗	✗	✓
MIME（我们的方法）	视频	✓	细粒度与粗粒度	✓

七种通用情绪类别：快乐、悲伤、中性、愤怒、厌恶、恐惧、以及惊讶。我们采用平衡重采样策略来缓解长尾类别偏差，从而确保对模型泛化能力进行严格评估。

通过模态退化技术，我们系统地将视频信号解耦为七个子集，以模拟现实世界中的不完整性：完整模态（FM）、面部细节缺失（FDM）、面部结构缺失（FSM）、视觉模态缺失（VMM）、面部细节和音频模态缺失（FDAM）、面部结构和音频模态缺失（FSAM）、音频模态缺失（AMM）。

3.2 情绪链（CoE）生成

为了赋予我们的数据集丰富且细粒度的推理标注，我们设计了一个全面的四阶段流程，利用**Qwen3- Omni** 生成高质量的情绪链。如图1所示，该流程明确引导模型在预测最终情绪之前，解耦并阐述其在各个模态中的观察结果。生成过程的详细信息见附录。CoE的逻辑流程分为三个部分：

- **场景理解**（C_{scene}）：刻画输入中存在的客观多模态线索，仅从可用模态（面部、身体、场景和音频）中提取事实信息。
- **情绪分析**（A_{emo}）：执行认知推理，将先前识别的线索映射到内部情绪状态。这一步明确阐述所观察信号如何导向情绪推断的逻辑路径。
- **预测情绪**（L）：基于前述系统分析，推导出与真实标签严格一致的最终离散情绪类别。

3.3 统计分析

为确保所提出基准的鲁棒性和泛化能力，我们对构建的 MIME 数据集进行了全面的统计分析，该数据集包含 2,000 个片段。图 2 和图 3 展示了 **MIME 统计概览**。如图所示，

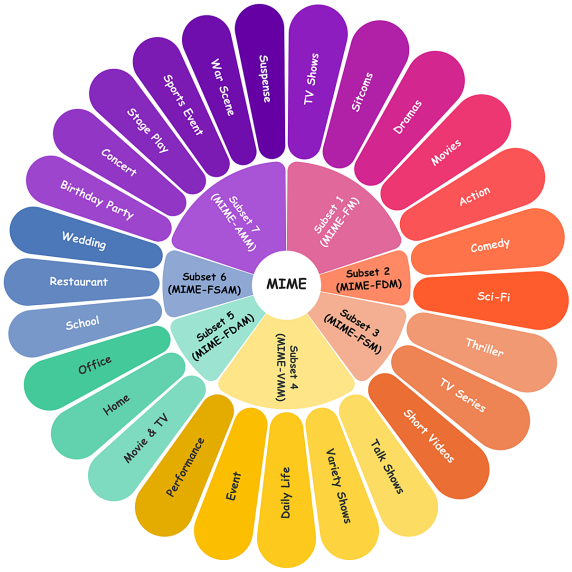

图 2：MIME 任务类型概览。不同视频场景下的任务类型。

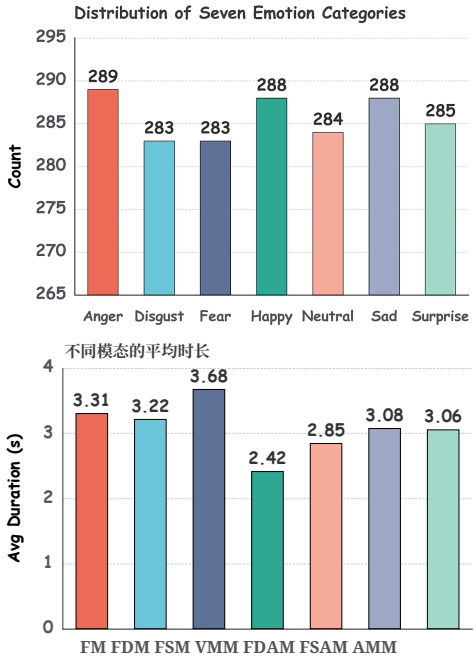

图 3：情绪类别（上）和各模态视频时长（下）的均衡分布。

任务类型（图 2）呈现出层次化分布：该数据集涵盖特定的模态缺失类型（内圈，例如 AudioMiss、严重退化的人脸、视觉缺失），并将其系统地解耦到多样化的真实场景上下文中（外圈，例如生活方式、电影、视频博客），从而实现细粒度的对多样化视频情境的分析。图 3 上方显示，数据集严格遵循七种通用情绪类别的均衡样本分布，有效缓解了野外数据集中普遍存在的长尾偏差。此外，如图 3 下方所示，不同模态组合下音频/视频数据的平均时长分布均匀，有助于全面评估各种退化情况下的时间理解能力。

3.4 评估套件

3.4.1 评估策略。传统评估框架主要依赖离散情绪标签，难以全面衡量模型在复杂的模态受损场景中的推理质量。受近期情绪智能基准中引入的免标注评估范式启发 [48]，，我们提出了一种以大语言模型裁判机制为基础的混合评估策略。具体而言，我们采用 Qwen3.5-plus 作为裁判模型，对受评模型生成的 CoE 进行自动化、细粒度的定量评分。该方法不仅评估最终分类准确率，还能客观衡量模型主观推理过程的逻辑严谨性和事实依据，同时避免高昂的人工标注成本。

3.4.2 评估指标。 为了从客观分类性能和主观可解释性两个维度全面剖析多模态大语言模型的情绪智能，我们将评估指标划分为**两个阶段**：

阶段 *I*：标准分类指标。 针对最终情绪预测，我们计算**准确率（Acc）**和**F1分数**，以严格评估模型在七种情绪类别上的分类鲁棒性，尤其是缓解模态退化情况下可能出现的长尾分布问题。

阶段 *II*：*CoE* 质量指标。 为了细致量化“补偿性推理”能力，大语言模型评审会依据四种预定义模态，对生成的 CoE 进行细粒度评分：面部（f）、身体（b）、场景（s）和音频（a）。我们将模态集合记为 $\mathrm{M} = \{f, b, s, a\}$，并令 $I_m \in \{0, 1\}$ 表示模态 m 是否可用于评分；若模态m 可用或虽已退化但仍可观察，则为 $I_m = 1$；若模态 m 确实缺失，则为 $I_m = 0$。评估包括三**个子指标**和一**个综合指标**：

上下文感知评分（CPS）：评估模型是否准确捕捉到残存的物理线索。总体 CPS 仅针对有效模态进行计算，其中 $s_m^{cps} \in [0, 10]$:

$$CPS = \frac{\sum_{m \in \mathcal{M}} I_m \cdot s_m^{cps}}{\sum_{m \in \mathcal{M}} I_m} \tag{1}$$

情感推断分数（AIS）：评估根据每种模态所感知线索推断情绪状态的逻辑严谨性，并明确对幻觉进行惩罚，其中每种有效模态的得分为 $s_m^{ais} \in [0, 10]$:

$$AIS = \frac{\sum_{m \in \mathcal{M}} I_m \cdot s_m^{ais}}{\sum_{m \in \mathcal{M}} I_m} \tag{2}$$

表 2：MIME 上的整体性能对比（%）。表现排名前三的结果分别以红色（1st）、蓝色（2nd）和黄色（3rd）标

模型	F1	CPS	AIS	LCS	HCI
Qwen3.5-27b [33]	40.99	6.57	5.13	4.32	5.46
Qwen3.5-plus [33]	42.60	**6.70**	**5.32**	4.43	**5.60**
gpt-5.4-nano [30]	26.21	5.41	3.97	2.93	4.23
gemini-3.1-flash [6]	**46.61**	6.54	5.24	**4.68**	5.59
gpt-5-mini [28]	33.52	5.91	4.42	3.48	4.73
Doubao-seed-2-0 [4]	45.27	6.43	4.96	4.65	5.46
GPT-5-nano [29]	27.01	5.53	4.06	2.95	4.31

标签一致性得分（LCS）：检查长文本推理路径的语义结论是否与最终类别预测严格一致。若正确，LCS 等于 10，否则为 0。

整体 CoE 指数 (HCI)：作为衡量模型综合“情感可解释性”的最终基准。它是上述三个指标的加权整合：

$$HCI = \alpha \cdot CPS + \beta \cdot AIS + (1 - \alpha - \beta) \cdot LCS \quad (3)$$

其中，超参数经经验设定为 $\alpha = 0.4$ 和 $\beta = 0.3$，以强调基础语境理解。

3.4.3 人工验证。为验证我们提出的多模态大语言模型评审框架的可靠性，我们邀请了 3 位专家标注员进行人工评估。专家标注了 173 个样本及其对应的 CoE 输出，以分析 Qwen 生成的分数与人工标注之间的一致性。如图 1 右下角面板所示，两种评估的斯皮尔曼秩相关系数为 [37] 0.9415（强正相关），科恩卡帕系数为 [27]0.8542（评定者间一致性近乎完美），组内相关系数（ICC）为 [14] 0.9628（评分一致性极佳）。这些指标证实，我们的多模态大语言模型评审方法与人工评估高度一致，验证了其作为评估方法的可信度和鲁棒性。

4 实验

4.1 实验设置

我们在 7 个前沿多模态大语言模型上评估性能，包括 GPT-5 系列（gpt-5.4-nano、gpt-5-mini、gpt-5-nano）、Gemini 系列（gemini-3.1-flash）、Qwen 系列（qwen3.5-plus、qwen3.5-27b）以及 doubao-seed-2-0。所有实验均在零样本设置下进行。此外，为实现可扩展的定性评估，我们采用 Qwen3.5-Plus 作为主要的大语言模型评审。借助结合多线程的并发执行架构，裁判模型依据生成的 CoE 文本在逻辑一致性和事实依据方面的表现进行评分。

4.2 主要结果

我们在表 2 和表 3 中全面展示了各种模型在 MIME 基准上的性能比较，涵盖标准分类指标（例如准确率和 WF1）以及由评审智能体评分的 CoE 质量指标。

总体分类与推理性能。实验结果清晰地揭示了在模态受损场景下进行多模态情感识别所面临的巨大挑战。即使是表现最佳的模型 gemini-3.1-flash，其 F1分数也仅为 46.61，HCI 为 5.59。紧随其后的是 doubao-seed-2-0（F1：45.27，HCI：5.46）和 qwen3.5-plus（F1：42.60，HCI：5.60）。相比之下，gpt-5.4-nano 和 gpt-5-nano 等较小规模模型的性能骤降，F1分数分别跌至 26.21 和 27.01，其 HCI 也仅徘徊在 4.23 和 4.31 左右。所有受评估模型的 F1 分数均低于 50，这表明即使是最先进的多模态大语言模型，在复杂的真实环境中发展稳健的情感智能方面仍处于初期阶段。

情境感知与情感推断之间的差距。深入分析 CoE 质量指标后，我们观察到“场景理解（CPS）”与“情绪分析（AIS）”之间存在显著差异。例如，qwen3.5-plus 在感知基础物理线索（CPS）方面得分高达 6.70，但在将这些线索转化为内在情感推断（AIS）时，得分却大幅降至 5.32。同样，gpt-5-mini 在这两项指标上的得分分别为 5.91 和 4.42。这一现象深刻揭示了当前多模态大语言模型的情感认知瓶颈：它们能够在退化环境中充分“看见”并提取客观的物理元素，却严重缺乏基于这些观察结果进行深入补偿性逻辑推理的能力。

不完整条件下的模态贡献。表3展示了不同模态对情感识别的影响。比较 FDM 与 FDAM 可见音频贡献：当面部细节退化时，移除音频会导致 WF1 明显下降，例如 qwen3.5-27b 从 46.81 降至 30.75（-16.06），gemini-3.1-flash 从 49.82 降至 37.31（-12.51）。这表明面部信息不可靠时，音频能够起到补偿作用。

4.3 观察与洞见

通过对定量分类结果和定性 CoE 评估进行交叉分析，我们提炼出关于 MLLMs 在模态受损条件下（子集 1–7）进行情感推理的三点洞见：

观察 *1*：视觉感知与认知推理之间存在显著差距。“看到”一个场景并不意味着“理解”其中的情绪。在所有模型中，上下文感知评分（CPS）始终高于情感推断分数（AIS）。例如，qwen3.5-plus 能准确提取物理线索（CPS：6.70/10），却难以在逻辑上推断内在情绪状态（AIS：5.32/10）。这表明，尽管多模态大语言模型具备扎实的视觉基础，却缺乏类似人类的共情认知，无法跨越从客观物理空间到主观情感空间的语义鸿沟。

观察*2*：稳健的隐式多模态情感识别需要跨模态补偿，而非单线索识别。子集级结果表明，不完整情感识别不能依赖单一主导线索。面部外观和音频提供直接的情感证据，

表3：不同子集上的分类性能。

模型	FM		FDM		FSM		VMM		FDAM		FSAM		AMM	
	准确率	WF1	准确率	WF1	准确率	WF1	准确率	WF1	准确率	WF1	准确率	WF1	准确率	WF1
gpt-5.4-nano [30]	22.81	18.92	25.95	19.70	31.12	27.80	43.50	42.28	20.67	16.29	31.94	29.72	25.00	21.03
gpt-5-mini [28]	27.07	24.13	24.68	22.80	36.93	34.70	**50.00**	**48.78**	23.08	21.98	44.27	42.36	31.75	31.64
gemini-3.1-flash [6]	46.10	44.83	**50.96**	**49.82**	**49.38**	**50.09**	44.25	42.75	37.98	37.31	48.42	47.33	**50.75**	**51.34**
gpt-5-nano [29]	23.75	19.87	20.89	16.48	28.22	25.22	49.50	47.32	19.23	16.61	35.94	33.47	21.75	17.98
qwen3.5-plus [33]	48.12	45.32	44.94	42.26	40.25	38.79	42.50	39.87	35.58	32.30	**50.26**	**49.28**	46.00	44.44
qwen3.5-27b [33]	44.11	40.57	48.73	46.81	42.74	40.79	42.61	39.48	35.10	30.75	46.35	44.43	43.50	40.64
doubao-seed-2-0 [4]	**51.13**	**48.10**	48.73	48.29	48.13	45.61	41.50	39.24	**40.87**	**38.02**	47.92	47.65	47.12	45.72

表4：不同模型的逐情绪准确率（ACC）和F1分数（F1）对比。

模型	快乐		Sad		中性		愤怒		厌恶		Fear		惊讶	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
gpt-5.4-nano [30]	15.28	0.23	26.04	0.27	33.10	0.28	53.50	0.41	9.19	0.16	29.72	0.35	8.42	0.14
gpt-5-mini [28]	21.18	0.28	33.68	0.36	51.41	0.31	57.14	0.43	11.66	0.20	45.23	0.45	23.51	0.30
gemini-3.1-flash [6]	50.69	0.57	53.31	**0.52**	68.20	**0.41**	56.14	0.52	**21.20**	**0.32**	37.81	0.46	40.14	**0.47**
gpt-5-nano [29]	12.15	0.19	24.31	0.29	**68.66**	0.31	35.07	0.39	4.59	0.09	**45.58**	0.38	16.49	0.24
qwen3.5-plus [33]	63.76	**0.59**	**56.25**	0.48	50.00	0.35	56.45	**0.54**	7.77	0.14	42.4	**0.48**	32.63	0.40
qwen3.5-27b [33]	66.67	0.58	55.56	0.45	43.66	0.35	50.87	0.49	6.71	0.12	42.05	0.45	35.92	0.43
doubao-seed-2-0 [4]	**68.75**	**0.59**	48.61	0.50	57.24	0.40	**57.84**	0.53	17.31	0.27	31.45	0.41	**43.51**	0.46

而身体姿势和场景上下文则充当补偿性线索。然而，当前的多模态大语言模型仍难以动态调整可用线索的权重。当面部细节和音频同时不可用时，模型主要根据肢体语言和场景上下文推断情感，导致性能明显下降。因此，IMER 的关键挑战在于：面对不同的缺失情况，自适应地在剩余模态之间转移注意力。

观察*3*：在复杂场景中，明确的*CoE*质量决定预测结果。我们的分析显示，分类准确率与CoE推理质量之间存在显著的正相关关系，该质量通过HCI进行衡量。最先进的模型（例如gemini-3.1-flash、qwen3.5-plus）

其HCI分数始终高于5.45，而轻量级GPT-nano模型的F1分数低于30，产生碎片化的推理轨迹，得分低于4.40。领先模型还获得了更高的标签一致性分数（LCS），这表明其预测依赖稳定的分步演绎，而非启发式捷径。因此，在多模态线索不完整或受损的情况下，高质量的CoE推理可作为可靠性锚点，使强化多模态大语言模型内部的“情感链补偿机制”成为缓解多模态情感瓶颈的有前景方向。



图 4：模型响应的错误案例。

5 结论

本文探讨了情感计算领域一个尚未得到充分研究的关键挑战：模态缺失条件下的多模态情感识别。我们进一步提出了MIME，这是首个细粒度的情感链基准。该基准系统地模拟了七种场景下的现实世界模态退化，以构建严格的测试环境。MIME通过自适应四阶段生成管线构建，并结合条件式幻觉抑制，生成高质量的三元结构化推理链。我们还设计了一个双阶段混合评估套件，采用大语言模型评审，量化模型客观的分类鲁棒性和主观的推理逻辑。

致谢

本研究得到中国国家自然科学基金（项目编号：62576085、62236006、62320106007 和 62521007）、中央高校基本科研业务费（项目编号：N25XQD057）以及中国国家博士后研究人员计划 C 类（项目编号：GZC20260869）的资助。本研究得到南京大学新型软件技术国家重点实验室开放研究基金（项目编号：KFKT2026B56）的资助。本研究得到湖北省多媒体与网络通信工程重点实验室开放课题的资助。本研究得到武汉大学国家多媒体软件工程技术研究中心的资助。本研究得到多模态人工智能系统国家重点实验室开放研究基金（项目编号：MAIS2026081）的资助。本研究得到辽宁省协同创新中心（CSLE）的资助。

参考文献

[1] Sree Bhattacharyya 和 James Z. Wang。2025。评估用于情感识别的视觉-语言模型。载于 计算语言学协会论文集：*NAACL 2025*年研究成果，Luis Chiruzzo、Alan Ritter 和 Lu Wang（编）。计算语言学协会，新墨西哥州阿尔伯克基，1798–1820。doi: 10 18653 1 2025 97 。/v / .findings-naacl。Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower、Samuel Kim、Jeannette N Chang、Sungbok Lee 和 Shrikanth S Narayanan。2008。IEMOCAP：交互式情绪双人动作捕捉数据库。语言资源与评测 42（2008），335–359。[3] Carlos Busso、Murtaza Bulut、Chi-Chun Lee、Abe Kazemzadeh、Emily Mower Provost、Samuel Kim、Jeannette Chang、Sungbok Lee 和 Shrikanth Narayanan。2008。IEMOCAP：交互式情绪双人动作捕捉数据库。语言资源与评测 42（12，2008），335–359。doi:10.1007/ s10579-008-9076-6[4] 字节跳动 Seed团队。2026。Seed2.0。https://seed.bytedance.com/en/seed2。字节跳动 Seed2.0模型官方页面。[5] Yuanyue Deng、Jintang Bian、Shisong Wu、Ji anhuang Lai 和 Xiaohua Xie。2025。用于无监督不完整多模态情感识别的多路图聚合与特征细化。信息融合 114（2025），102711。[6] Google。2026。Gemini 3开发者指南。https://ai.google.dev/gemini-api/docs/ gemini-3。Google AI官方开发者文档。[7] Zirun Guo、Tao Jin 和 Zhou Zhao。2024。利用模态缺失进行情感分析与情感识别的多模态提示学习。arXiv: 2407 05374 [ ] 2407 05374 。cs.CL https://arxiv.org/abs/ 。Zirun Guo、Tao Jin 和 Zhou Zhao。。利用模态缺失进行情感分析与情感识别的多模态提示学习。arXiv:2407.05374 [cs.CL] https://arxiv.org/abs/2407.05374[9] Guimin Hu、Weimin Lyu、Chang Sun、Zhihong Zhu、Lin Gui、Ruichu Cai、Erik Cambria 和 Hasti Seifi。2026。多模态情感计算的最新进展：自然语言处理视角。arXiv:2409.07388 [cs.CL] https://arxiv.org/abs/2409。07388[10] He Hu、Lianzhong You、Hongbo Xu、Qianning Wang、Fei Richard Yu、Fei Ma、Zebang Cheng、Zheng Lian、Yucheng Zhou 和 Laizhong Cui。2026。EmoBench-M：评测多模态大语言模型的情绪智能。arXiv:2502.04424 [cs.CL] https://arxiv.org/abs/2502.04424[11] Xingxun Jiang、Yuan Zong、Wenming Zheng、Chuangao Tang、Wanchuang Xia、Cheng Lu 和 Jiateng Liu。2020。DFEW：用于野外环境动态面部表情识别的大规模数据库。arXiv:2008.05924 [cs.CV] https://arxiv.org/ abs/2008.05924[12] Yuehan Jin、Xiaoqing Liu、Yiyuan Yang、Zhiwen Yu、Tong Zhang 和 Kaixiang Yang。2025。RoHyDR：用于不完整多模态情感识别的鲁棒混合扩散恢复。arXiv:2505.17501 [cs.CV] https://arxiv.org/abs/ 2505.17501[13] Sepideh Kalateh、Luis A Estrada-Jimenez、Sanaz Nikghadam-Hojjati 和 Jose Barata。2024。多模态情感识别的系统综述：构成模块、当前状态、应用与挑战。*IEEEAccess* 12（2024），103976–104019。[14] Gary G Koch。2004。组内相关系数。统计科学百科全书 （2004）。[15] Jiyoung Lee、Seungryong Kim、Sunok Kim、Jungin Park 和 Kwanghoon Sohn。2019。上下文感知情感识别网络。arXiv:1908.05913 [cs.CV] https://arxiv.org/abs/ 1908 05913 。Huihui Li、Huiqi Han、Chunlin Xu、Tongbao Chen、Xiaoyong Liu 和 Guihua Wen。2026。从完整模态到缺失基于文本、音频和视觉模态：综述。《人工智能工程应用》170（2026），114127。doi:10.1016/j.engappai.2026.114127[17] Mingcheng Li、Dingkang Yang、Xiao Zhao、Shuaibing Wang、Yan Wang、Kun Yang、Mingyang Sun、Dongliang Kou、Ziyun Qian 和 Lihua Zhang。2024。具有不完整模态的多模态情感分析的相关性解耦知识蒸馏。arXiv:2404.16456 [cs.CV] https://arxiv.org/abs/2404.16456[18] Zheng Lian、Rui Liu、Kele Xu、Bin Liu、Xuefei Liu、Yazhou Zhang、Xin Liu、Yong Li、Zebang Cheng、Haolin Zuo、Ziyang Ma、Xiaojiang Peng、Xie Chen、Ya Li、Erik Cambria、Guoying Zhao、Björn W. Schuller 和 Jianhua Tao。2025。MER 2025：当情感计算遇上大语言模型。载于第33届 ACM国际多媒体会议论文集（爱尔兰都柏林）（MM '25）。计算机协会，纽约州纽约市，美国，13837–13842。doi:10.1145/3746027.3762007[19] Zheng Lian、Rui Liu、Kele Xu、Bin Liu、Xuefei Liu、Yazhou Zhang、Xin Liu、Yong Li、Zebang Cheng、Haolin Zuo、Ziyang Ma、Xiaojiang Peng、Xie Chen、Ya Li、Erik Cambria、Guoying Zhao、Björn W. Schuller 和 Jianhua Tao。2025。MER 2025：当情感计算遇上大语言模型。arXiv:2504.19423 [cs.HC] https://arxiv.org/abs/2504.19423[20] Zheng Lian、Haiyang Sun、Licai Sun、Haoyu Chen、Lan Chen、Hao Gu、Zhuofan Wen、Shun Chen、Siyuan Zhang、Hailiang Yao、Bin Liu、Rui Liu、Shan Liang、Ya Li、Jiangyan Yi 和 Jianhua Tao。2025。OV-MER：迈向开放词汇多模态情绪识别。arXiv:2410.01495 [cs.HC] https://arxiv.org/ abs/2410.01495[21] Zheng Lian、Haiyang Sun、Licai Sun、Jiangyan Yi、Bin Liu 和 Jianhua Tao。2024。AffectGPT：用于可解释多模态情绪识别的数据集与框架。arXiv:2407.07653 [cs.HC] https://arxiv.org/abs/2407.07653[22] Zheng Lian、Licai Sun、Yong Ren、Hao Gu、Haiyang Sun、Lan Chen、Bin Liu 和 Jianhua Tao。2024。MERBench：多模态情绪识别统一评测基准。arXiv: 2401 03429 [ ] 2401 03429 . cs.HC [23] 2023 https://arxiv.org/abs/ 。Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。。视觉指令调优。arXiv: 2304 08485 [ ] 2304 08485 . cs.CV [24] https://arxiv.org/abs/ 。Rui Liu、Haolin Zuo、Zheng Lian、Björn W Schuller 和 Haizhou Li。2024。基于对比学习的模态不变特征获取：面向模态缺失的鲁棒多模态情绪识别。IEEE情感计算汇刊 15，4（2024），1856–1873。[25] Rui Liu、Haolin Zuo、Zheng Lian、Xiaofen Xing、Björn W. Schuller 和 Haizhou Li。2024。多模态对话中的情绪与意图联合理解：一个基准数据集。arXiv:2407.02751 [cs.CL] https://arxiv.org/abs/2407.02751[26] Yuanyuan Liu、Wei Dai、Chuanxu Feng、Wenbin Wang、Guanghao Yin、Jiabei Zeng 和 Shiguang Shan。2023。MAFW：用于自然场景动态面部表情识别的大规模多模态复合情感数据库。arXiv: 2208 00847 [ ] 2208 00847 . cs.CV [27] 2012 https://arxiv.org/abs/ 。Mary L McHugh。。评定者间信度：κ统计量。《Biochemia medica》22，3（2012），276–282。[28] OpenAI。2026。GPT-5 mini模型。https://developers.openai.com/api/docs/ models/gpt-5-mini。OpenAI官方 API 模型文档。[29] OpenAI。2026。GPT-5 nano模型。https://developers.openai.com/api/docs/ models/gpt-5-nano。OpenAI官方 API 模型文档。[30] OpenAI。2026。推出 GPT-5.4 mini 和 nano。https://openai.com/index/ introducing-gpt-5-4-mini-and-nano/。OpenAI官方发布页面。[31] Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea。2019。MELD：用于对话中的情绪识别的多模态多方数据集。arXiv:1810.02508 [cs.CL] https: // arxiv.org/abs/1810.02508[32] Xihang Qiu、Jiarong Cheng、Yuhao Fang、Wanpeng Zhang、Yao Lu、Ye Zhang 和 Chun Li。2025。不完整模态下用于情绪识别的联邦对话语义扩散。arXiv:2511.00344 [cs.CV] https://arxiv.org/abs/ 2511.00344[33] Qwen团队。2026。Qwen3.5：迈向原生多模态智能体。https://qwen. ai/blog?id=qwen3.5。Qwen官方发布博客。[34] Sahand Sabour、Siyang Liu、Zheyuan Zhang、June M. Liu、Jinfeng Zhou、Alvionna S. Sunaryo、Juanzi Li、Tatia M. C. Lee、Rada Mihalcea 和 Minlie Huang。2024。EmoBench：评估大语言模型的情绪智能。arXiv: 2402 12071 [ ] 2402 12071 . cs.CL [35] 2025 https://arxiv.org/abs/ 。Yuntao Shou、Tao Meng、Wei Ai 和 Keqin Li。。多模态大语言模型遇上多模态情绪识别与推理：综述。arXiv:2509.24322 [cs.CL] https://arxiv.org/abs/2509.24322[36] Shezheng Song、Chengxiang He、Shan Zhao、Chengyu Wang、Qian Wan、Tianwei Yan 和 Meng Wang。2025。MOSABench：用于评估多模态大语言模型对复杂图像理解能力的多对象情感分析基准。arXiv: 2412 00060 [ ] 2412 00060 . cs.CV [37] 1904 https://arxiv.org/abs/ 。C. Spearman。。《两事物间关联的证明与测量》。《The American Journal of Psychology》15，1（1904），72–101。http://www. jstor.org/stable/1412159[38] Haoqin Sun、Shiwan Zhao、Shaokai Li、Xiangyu Kong、Xuechen Wang、Aobo Kong、Jiaming Zhou、Yong Chen、Wenjia Zeng 和 Yong Qin。2024。增强

不完整数据中的情绪识别：一种新颖的跨模态对齐、重建与精炼框架。arXiv:2407.09029 [cs.MM] https: //arxiv.org/abs/2407.09029[39] Jen tse Huang、Man Ho Lam、Eric John Li、Shujie Ren、Wenxuan Wang、Wenxiang Jiao、Zhaopeng Tu 和 Michael R. Lyu。2024。情感麻木还是富有同理心？使用 EmotionBench 评估大语言模型的情感体验。arXiv:2308.03656 [cs.CL] https://arxiv.org/abs/ 2308 03656 。Paul Waligora、Haseeb Aslam、Osama Zeeshan、Soufiane Belharbi、Alessandro Lameiras Koerich、Marco Pedersoli、Simon Bacon 和 Eric Granger。2024。自然场景中的情绪识别联合多模态变换器。arXiv:2403.10488 [cs.CV] https://arxiv.org/abs/2403.10488[41] Yan Wang、Yixuan Sun、Yiwen Huang、Zhongying Liu、Shuyong Gao、Wei Zhang、Weifeng Ge 和 Wenqiang Zhang。2022。FERV39k：用于视频中面部表情识别的大规模多场景数据集。arXiv:2203.09463 [cs.CV] https://arxiv.org/abs/2203.09463[42] ChengYan Wu、Yiqiang Cai、Yang Liu、Pengxu Zhu、Yun Xue、Ziwei Gong、Julia Hirschberg 和 Bolei Ma。2025。对话中的多模态情绪识别：方法、趋势、挑战与展望综述。载于 计算语言学协会：*EMNLP 2025*论文集，Christos Christodoulopoulos、Tanmoy Chakraborty、Carolyn Rose 和 Violet Peng（编）。计算语言学协会，中国苏州，6257–6274。doi:10.18653/v1/2025.findings-emnlp.332[43] Chengyan Wu、Yiqiang Cai、Yang Liu、Pengxu Zhu、Yun Xue、Ziwei Gong、Julia Hirschberg 和 Bolei Ma。2025。对话中的多模态情绪识别：方法、趋势、挑战与展望综述。arXiv:2505.20511 [cs.CL] https://arxiv.org/abs/2505.20511

[44] Renjie Wu、Hu Wang、Hsiang-Ting Chen 和 Gustavo Carneiro。2026。《模态缺失条件下的深度多模态学习：综述》。arXiv:2409.07825 [cs.CV] https://arxiv.org/abs/2409.07825[45] Xiaocui Yang、Wenfang Wu、Shi Feng、Ming Wang、Daling Wang、Yang Li、Qi Sun、Yifei Zhang、Xiaoming Fu 和 Soujanya Poria。2025。《MM-InstructEval：多模态大语言模型在多模态推理任务上的零样本评估》。arXiv: 2405 07229 [ ] 2405 07229 。[46] cs.MM https://arxiv.org/abs/ 。Adnan Yazici、Tayfun Kucukyilmaz、Tansel Dokeroglu、Aidana Sharipbay、Min-Ho Lee 和 Benjamin Tyler。2026。《最先进的多模态情绪识别：全面综述与分类体系》。《智能系统与应用》30 (2026 200642 10 1016 2026 200642)，。doi: 。/j.iswa. 。Amir Zadeh、[47] Rowan Zellers、Eli Pincus 和 Louis-Philippe Morency。2016。《MOSI：在线观点视频中的情感强度与主观性分析多模态语料库》。arXiv:1606.06259 [cs.CL] https://arxiv.org/abs/1606.06259[48] Fan Zhang、Zebang Cheng、Chong Deng、Haoxuan Li、Zheng Lian、Qian Chen、Huadai Liu、Wen Wang、Yi-Fan Zhang、Renrui Zhang、Ziyu Guo、Zhihong Zhu、Hao Wu、Haixin Wang、Yefeng Zheng、Xiaojiang Peng、Xian Wu、Kun Wang、Xiangang Li、Jieping Ye 和 Pheng-Ann Heng。2026。《MME-Emotion：多模态大语言模型情绪智能的整体评估基准》。arXiv: 2508 09210 [ ] 2508 09210 。cs.CV [49] https://arxiv.org/abs/ 。Guowei Zhong、Ruohong Huan、Mingzhen Wu、Ronghua Liang 和 Peng Chen。2025。《面向模态缺失与分布偏移条件下的鲁棒多模态情绪识别》。arXiv:2506.10452 [cs.CV] https://arxiv.org/abs/2506. 10452[50] Youran Zhou、Sunil Aryal 和 Mohamed Reda Bouadjenek。2024。《使用特殊缺失机制处理缺失数据的综述》。arXiv:2404.04905 [stat.ME] https://arxiv.org/abs/2404.04905