/TemplateVersion (2026.1)

HumanSense：从多模态感知到通过推理的多模态大模型实现同理心的上下文感知响应

Zheng Qin[1,2], Ruobing Zheng[2†], Yabing Wang[1], Tianqi Li[2], Yi Yuan[2], Jingdong Chen[2], Le Wang[1*]

Abstract

虽然多模态大语言模型（MLLMs）显示出实现真正像人类一样的互动的巨大潜力，但由于缺乏细粒度的评估框架来针对以人为中心的场景进行评估，进展受到了阻碍，这些场景包括对复杂人类意图的理解以及提供具有同情心的、符合情境的回应。在此，我们介绍 HumanSense，一个旨在评估 MLLMs 在人性化感知和互动能力方面的综合基准，特别强调对扩展多模态情境的深刻理解和合理反馈的形成。我们的评估表明，领先的 MLLMs 在提高方面仍有很大空间，特别是在高级互动导向的任务中。通过视觉输入补充音频和文本信息获得了显著提升，而全模态模型在这些任务中表现出优势。此外，我们认为适当的反馈来自于对对话者需求和情感的背景分析，推理能力是实现这一目标的关键。据此，我们采用多阶段的、模态渐进的强化学习来增强一个全模模型的推理能力，在评估结果上取得了显著提升。此外，我们观察到成功的推理过程展现了高度一致的思维模式。通过设计对应的提示，我们还以无训练方式增强了非推理模型的性能。项目页面：https://digital-avatar.github.io/ai/HumanSense/

介绍

科幻小说 (??) 经常描绘一个未来，其中人工智能不仅仅作为任务执行的工具，还作为提供社会支持和情感联系的人类伴侣。基础的演变从狭窄、以任务为导向的系统到通用人工智能的过程是以能够通过语言、表情和肢体语言理解人类意图为前提，从而能够做出适当的回应。

多模态大语言模型（MLLMs） (????) 为实现这一愿景提供了一条有前途的途径。它们能够整体处理视觉、听觉和文本信息，从而对用户和环境进行全面的理解。MLLMs 还具备深入分析感知信息 (?) 的潜力，并随后规划适当的反馈，这不仅限于文本响应，还可以包括在时间序列中合适的情感、语气和手势标签。这样的输出可以进一步与视频生成 (??)、语音合成 (??) 和会说话的头像 (????) 方法集成，以提供高度拟人化的交互体验。

实现这一目标首先需要定义必要的能力，评估模型性能，然后应用优化。然而，现有的基准 (???) 缺乏针对这些以人为中心的场景进行的目标明确、细粒度的评估。为了解决这一差距，我们首先定义了 MLLMs 在这种场景中所需的主要能力：1）多模态感知，2）隐含信息的上下文理解，和 3）在多轮交互中给出适当响应。对于互动场景，我们同时考虑响应内容和响应策略。

基于上述考虑，我们提出了 HumanSense 基准测试。该基准测试包括 15 项逐步挑战的测试，总计 3,882 道问题，这些问题均来自真实世界的记录。在交互测试中，多模态大语言模型（MLLMs）被要求在互动中扮演一方角色并生成响应，然后与真实的人类记录进行比较。我们对当前领先的 MLLMs 进行了全面评估，包括视觉-语言模型 (??)、Omni 模型 (??)、音频-语言模型 (?)。结果显示在人类中心的场景中，特别是高级互动导向任务中，仍有显著的改进空间。模态消融研究表明，视觉、听觉和文本信息在高级任务中均起关键作用，而能够联合处理音频、视频和文本的 omni 模型则展现出明显优势。

基于这些发现，我们提出全模态推理可以增强 MLLMs 的认知和交互能力。这个主张源于我们观察到，在交流中适当的反馈依赖于对全模态上下文、对话者的需求、情感和个人特征的全面考虑。因此，我们采用多阶段、全模态的强化学习方法来构建推理全模态模型，从而在评估中取得了显著的改进。此外，我们观察到成功的推理过程展示出高度一致的模式。通过设计相应的提示，我们还以无需训练的方式提高了非推理模型的性能。

通往通用人工智能的道路需要长期的、多方面的探索。通过这项工作，我们旨在激励社区探索 Omni MLLMs 在改善以人为中心的 AI 交互方面的潜力，并帮助塑造这一新兴方向。

相关工作

多模态大型语言模型。 LLMs (?????) 已被广泛用于人类行为和情感分析，促进了包括对话模拟、行为预测和文本情感分类的应用。这些能力支持从社交媒体监控到自动客户支持系统等多种场景。然而，仅凭文本信息常常不够，因为这些模型缺乏对视觉线索的支持，比如面部表情和肢体语言，这些对于全面的人类行为分析至关重要。视觉 MLLMs (??????) 显示出强大的视觉理解能力，可以通过面部表情和肢体语言的分析准确识别情绪和行为。然而，这些模型的一个关键限制是无法处理音频信息，导致对话内容、声调和环境声音等重要听觉线索的缺失。这一限制在其对复杂现实场景的理解中造成了显著的差距并引入了偏差。相比之下，Omni 模

† Co-first author. Project lead. * Corresponding author.
[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University. [2] Ant Group.

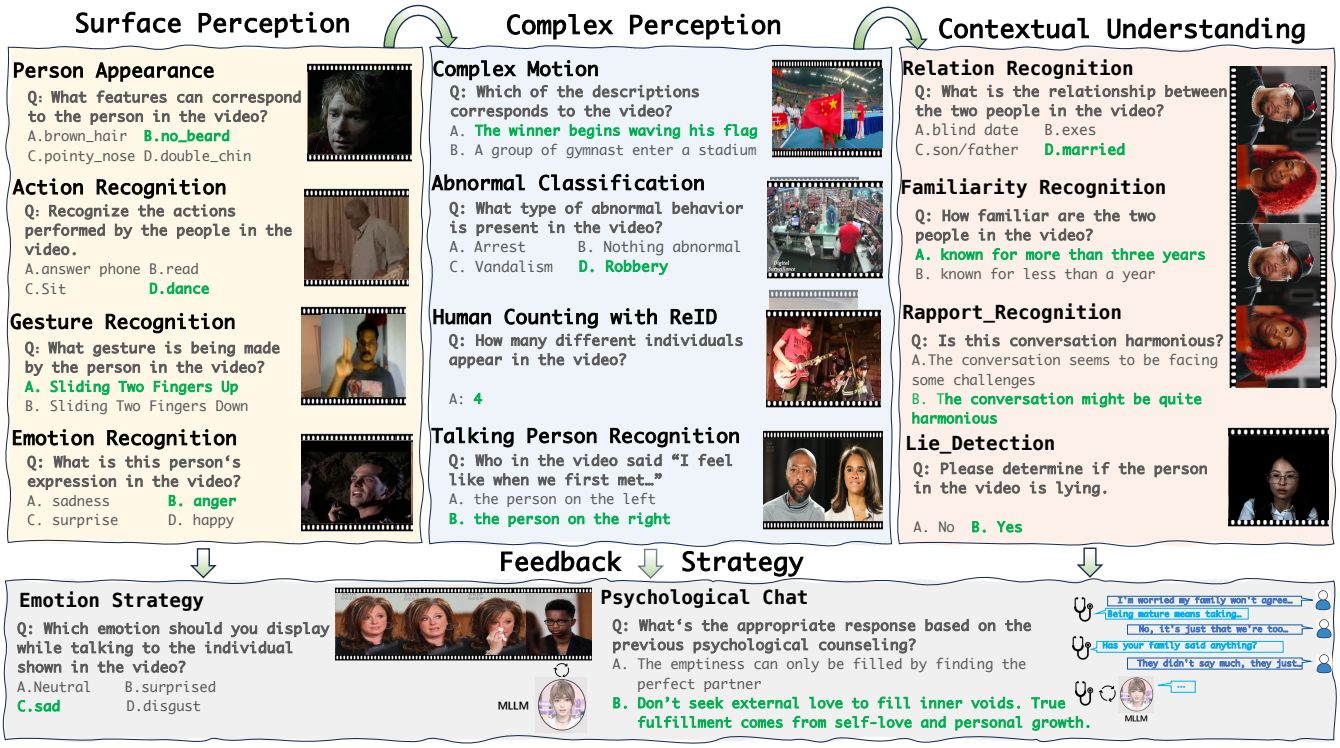

Figure 1: HumanSense 中任务的演示。注意：为了清晰和简洁，上述问题稍作简化。

型 (???????) 整合了包括视觉、语言和音频在内的多种模态，以提供对复杂人类互动的全面模拟。这些模型可以在处理对话内容的同时分析视觉线索，从而实现对人类交流动态的更细致和准确的理解。

随着多模态大模型的进步，已经出现了几个评价基准，其中大多数集中在评估视频理解能力上。此外，StreamingBench 专注于流媒体视频理解。然而，评估大模型从以人为中心的角度出发的基准很少，这对于此类模型在实际场景中的应用至关重要。HumanOmniV2 专注于全模态方式解读意图、解释情感和检测潜在的欺骗行为。虽然 HumanOmniV2 在以人为中心的视频理解上提供了有价值的见解，但它缺乏对响应规划或互动能力的评估。

人类感知基准

概述

我们旨在通过 HumanSense 框架系统地评估 MLLM 的人本能力：1）以人为中心的多模态感知，2）隐含信息的情境理解，以及 3）交互场景中的响应策略，如图 1 。评估任务根据难度递增被组织成一个四层金字塔结构 (L1–L4)，如图 ?? （左）所示：

- 感知（L1 & L2）：基础层集中于 L1 的单模态、表面感知任务，而 L2 处理多模态、长时间复杂感知任务。二者结合，共同形成智能的基础能力。
- 理解（L3）：在感知基础上，这一层评估模型是否能够揭示对话中隐含的信息。
- 响应（L4）：作为能力的顶峰，此层评估模型在各种互动场景中生成适当且合理响应的能力。

这个设计确保了评估的系统性：模型必须首先具备坚实的感知基础，才能逐步达到深刻理解，最终在顶层做出明智的反馈决策。请参阅图 ?? 获取 HumanSense 任务概述以及关于任务数量分布和视频长度的数据集统计信息。

任务定义

在人类交流中，多样的信息通过不同的方式传播。例如，视觉表达，如面部表情和手势，可以传递情感或语义信息；声音可以直接表达内容或间接传递情感。感知这些基本信息对互动至关重要。我们设计了以下多模态感知任务。

L1：**表面感知**

- 人物外观（PA）评估模型对面部外观的细粒度感知，因为外观构成了人物识别的基本方面。我们利用从 CelebV-HQ (?) 的数据注释来设计一系列多项选择题。每个问题询问视频中的人物是否具有特定属性，例如“男性”、“年轻”、“丰满”、“红润的脸颊”、“椭圆形脸”或“直发”。
- 动作识别（AR）旨在评估模型识别基本人类动作的能力。我们要求 MLLMs 从来自 AVA (Atomic Visual Actions) (?) 数据集的电影剪辑中识别当前动作。这些动作包括个人行为（例如，“走路”、“睡觉”）、与物体的互动（例如，“打开窗户”、“划船”），以及人际互动（例如，“吻一个人”、“与某人交谈”）。
- 手势识别（GR）旨在评估模型识别手势的能力，这些手势在交流中传达了丰富的语义信息。我们使用 Jester (?) 数据集构建单项选择问题，涉及到的手势包括“向前滚动手”和“两指向上滑动”。

- 情感识别（ER）研究对面部表情的识别，因为它们是表达情感的主要方式。基于 CelebV-HQ （?）数据集，我们使用其内置标签生成单选题，要求模型识别视频中的情感，如“快乐”、“悲伤”、“厌恶”、“愤怒”等。

L2: **复杂感知**

- 复杂运动（CM）研究扩展复杂动作序列的描述，其与目标行为的理解相关。我们利用来自 ActivityNet 数据集 （?）的字幕来构建单项选择题，其中模型必须识别出在长视频片段中执行的动作的正确描述。
- 异常分类（AC）评估异常人类行为的检测。我们根据 UCF-Crime100 数据集制定单项选择题 （?）。异常事件包括“偷窃”、“抢劫”或“打斗”等。
- 通过 ReID 进行人计数 (HC) 评估模型识别和记住个体的能力。我们使用跟踪数据集 TAO （?）来计算个体数量，并询问模型整个视频中出现的不同人物的总数。一些具有挑战性的问题涉及摄像机转换以及人类的间歇出现和消失。
- 说话人识别（TR）评估通过整合视觉和听觉信息进行判断的能力。基于 RealTalk 数据集 （?），我们提取视频片段并制定单选问题，要求模型识别出哪个人正在说特定内容。

L1 和 L2 任务评估多模式大模型在“看”和“听”方面的感知能力。实现和谐的交流需要深入思考上下文内容，并提供与对话者情感相对应的适当回应。因此，L3 检查模型在交互中“理解”隐含信息的能力，而 L4 评估模型在不同场景中的“回应”能力。

L3：**上下文理解**

- 熟悉度识别（FR1）评估模型理解人际互动和感知人际亲密度的能力。基于《The Skin Deep》频道中对话参与者之间的相识时间，我们构建单选题，让模型判断视频中个体之间的熟悉程度。
- 关系识别 (RR1) 评估模型感知交流氛围是否和谐的能力。我们从 The Skin Deep 频道中标注视频聊天内容，涵盖多个维度，包括互动频率、交流氛围以及观点冲突程度，以构建单选题。评估包括互动氛围是否愉快以及谈话内容中是否存在冲突。
- 关系识别（RR2）评估模型通过多模态上下文判定人类关系的能力。基于“The Skin Deep”频道中对话参与者之间的关系，我们构建单项选择题，要求模型预测视频中个体之间的关系类型，如“已婚”、“兄弟姐妹”、“儿子/母亲”等。模型必须整合视觉外观、年龄差异和对话内容的信息来做出判断。
- 测谎（LD）评估模型检测人类撒谎的能力。我们基于 SEUMLD 数据集制定了单选题。模型需要通过分析视觉和音频线索来判断每个视频片段中的说话者是否在撒谎。
- 诈骗识别（FR2）通过大约一分钟的电话录音评估特定诈骗行为的检测，这些录音取自电信诈骗文本数据集 (?)。诈骗类型包括“贷款和信用卡代理诈骗”、“冒充公安、司法和政府机构”以及“冒充领导或熟人”等。

第四部分：反馈策略

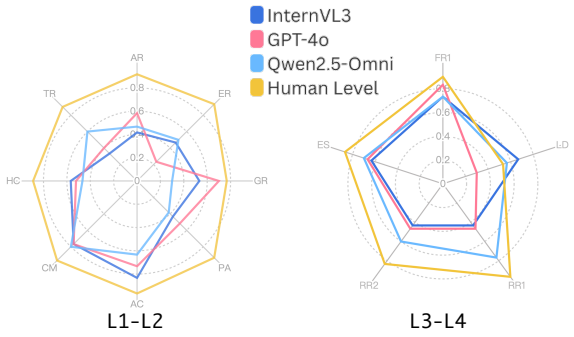

Figure 2: HumanSense（mini）上的性能雷达图。结果包括具有多种最先进的多模态模型的人类水平性能。

- 情感策略（ES）评估多模态大模型（MLLM）在交流过程中提供适当面部表情反馈的能力，创建一种有同理心的互动体验。基于 RealTalk 数据集 （?），我们提取视频片段，将说话者用作问题的输入，并注释倾听者的面部表情作为答案以构建单项选择题。
- 心理聊天（PC）评估 MLLMs 在复杂、长情景交互中生成适当回应的能力。在此，我们使用了专业的在线心理学数据集，情感急救数据集 （?），从中我们构建了单项选择题。模型需要基于之前的多轮对话选择一个适当的回应。

数据构建

问答生成 HumanSense 包括 3,291 个基于视频的问题和 591 个基于音频的问题。相关数据源自现有的开源数据集和 YouTube 视频。我们使用模板构建问题-答案（QA）对，利用现有数据集中的注释和内置标签。我们还使用各种现成的模块，包括情感识别、大型语言模型（LLMs）和光学字符识别（OCR），来分析源数据并提取与任务相关的信息。每个任务的构建详细信息在补充材料中提供。

问答增强 为了提高评估的泛化能力并避免评估偏差，我们对问题和答案进行增强。对于每个问题，我们设计了不同的候选模板以供随机选择。对于答案，我们平衡正确选项的分布并使正确和错误的选择长度相等。这种问答增强也避免了在后续强化学习实验中出现奖励作弊（?）。

我们对所有问答对的 20% 样本进行人工检查，特别关注 L3 和 L4 任务。例如，在“心理对话”中，我们确保内容包含足够的多轮对话。我们还严格验证正确选项，以确保它们提供专业和适当的建议。在“互动关系识别”任务中，我们设计了评分维度，包括互动频率、交流氛围和内容和谐性，用于 LLM 评估，随后进行结果的人工质量检查。

关于 HumanSense 的评价

我们对领先的多模态大型语言模型 (MLLMs) 进行了全面评估，模型规模最多达 10B，包括：(1) 视觉 LLMs，代表当今 MLLMs 的最主流分支；(2) 音频 LLMs；以及 (3) 本地设计用于整合视觉、音频和文本的全模态 LLMs。对于视觉 LLMs，我们评估了模型，如 Qwen2.5-VL （?）、

Models	Avg.	AR	ER	GR	PA	AC	CM	HC	TR	FR1	FR2	LD	RR1	RR2	ES	PC
		L1				L2				L3					L4	
	HumanSense (tiny) Perf.															
Human Level †	0.875	0.917	0.933	0.767	0.933	0.967	0.967	0.889	0.900	0.900	0.800	0.533	0.967	0.833	0.867	0.933
GPT-4o †	0.552	0.583	0.233	0.700	0.517	0.733	0.767	0.522	0.400	0.833	-	0.300	0.467	0.467	0.667	-
InternVL3-8B †	0.558	0.417	0.467	0.533	0.433	0.833	0.767	0.567	0.333	0.733	-	0.667	0.433	0.433	0.633	-
Qwen2.5-Omni-7B †	0.578	0.467	0.500	0.300	0.383	0.633	0.800	0.467	0.600	0.733	0.700	0.567	0.767	0.600	0.700	0.367
Qwen2-Audio-7B †	-	-	-	-	-	-	-	-	-	-	0.333	-	-	-	-	0.333
	HumanSense Perf.															
Proprietary Models (API)																
GPT-4o	0.557	0.548	0.282	0.620	0.517	0.750	0.776	0.536	0.570	0.735	-	0.310	0.480	0.535	0.587	-
Vl-Model																
LLaVA-Next-Video-7B	0.479	0.500	0.480	0.263	0.583	0.440	0.413	0.264	0.487	0.665	-	0.505	0.560	0.500	0.577	-
Qwen2-VL-7B	0.507	0.473	0.470	0.307	0.322	0.600	0.591	0.424	0.537	0.665	-	0.515	0.627	0.495	0.570	-
Qwen2.5-VL-7B	0.512	0.540	0.497	0.267	0.448	0.550	0.644	0.461	0.523	0.480	-	0.545	0.627	0.485	0.590	-
VideoLLaMA3-7B	0.520	0.543	0.463	0.323	0.517	0.530	0.694	0.561	0.493	0.610	-	0.515	0.587	0.400	0.530	-
LLaVA-OneVision-7B	0.521	0.545	0.510	0.400	0.592	0.620	0.676	0.268	0.503	0.600	-	0.530	0.560	0.430	0.543	-
InternVL3-8B	0.561	0.393	0.483	0.387	0.547	0.670	0.751	0.630	0.567	0.735	-	0.555	0.527	0.490	0.557	-
Audio-Model																
Qwen2-Audio-7B	-	-	-	-	-	-	-	-	-	-	0.437	-	-	-	-	0.399
Omni-Model																
Ola-7B	0.525	0.557	0.463	0.263	0.573	0.320	0.420	0.371	0.597	0.785	0.733	0.565	0.653	0.640	0.577	0.567
IXC2.5-OmniLive-7B	0.544	0.508	0.467	0.257	0.338	0.660	0.584	0.544	0.533	0.780	0.415	0.500	0.560	0.470	0.463	0.324
Qwen2.5-Omni-7B	0.554	0.473	0.513	0.303	0.350	0.600	0.630	0.438	0.600	0.770	0.740	0.550	0.713	0.650	0.607	0.399

Table 1: 在 HumanSense 上的评估。† 表示 HumanSense（微型）集上的结果，用于与人类水平表现进行比较。虽然 GPT-4o 被设计为多模态，但由于当前 API 不支持音频输入，无法对音频相关任务（FR2, PC）进行评估。为了在所有模型中进行公平比较，多模态的总体平均分（Avg.）是在去除两个音频任务的结果后计算的。每个指标的最佳结果已加粗显示。

Qwen2-VL (?)、LLaVA-OneVision (?)、LLaVANeXT-Video (?)、VideoLLaMA3 (?) 和 InternVL3 (?)。在全模态模型中，我们测试了 Qwen2.5-Omni (?)、IXC2.5OmniLive (?) 和 Ola (?)。对于音频 LLMs，我们评估了 Qwen2-Audio (?)。此外，我们测试了强大的全模态模型 GPT-4o (?)。所有评估都在零样本设置中进行，使用每个模型提供的默认提示。对于视频处理，我们遵循官方配置，涵盖了关键参数，包括帧提取方法、每秒帧数（FPS）和分辨率。我们将向社区发布我们的数据和代码，以促进对各类模型的更广泛评估。

为了适应每个模型的特性，视觉 LLM 仅在视频任务上进行评估。同样，音频 LLM 也仅在音频任务上进行评估。相比之下，旨在进行多模态处理的全能模型需要完成所有任务。

为了建立一个人类表现的基准，我们策划了一个新的评估集，命名为 HumanSense (tiny)，包括在我们的任务中随机抽取的 450 个问题（每个任务 30 个问题）。我们首先通过让评估者独立回答每个问题来建立一个人类基线。随后，我们将几个领先模型的表现与人类标准进行比较，包括 GPT-4o (?)，Intern3-VL (?)，Qwen2-Audio (?) 和 Qwen2.5-Omni (?)。

结果

人在我们的基准测试中平均达到 87.5% 的准确率，超出表现最佳的模型 29.7%。如图 2 所示，所有模型与人类水平的性能之间仍存在较大的差距，尤其是在复杂的 L3-L4 任务中，强调了当前 MLLM 在以人为中心的任务上能力需要显著提升。

如表 1 所示，Intern3-VL 在平均性能方面显示出最大的优势，在 L1-L2 感知任务和 L3-L4 高级任务中均具有出色表现。具体而言，LLaVA-OneVision 在 L1 系列任务中取得了最高的指标，反映了其在基础任务上的卓越视觉感知能力。InternVL3 在 L2 和一些 L3 任务中表现突出，展示了强大的长视频记忆和上下文理解，这

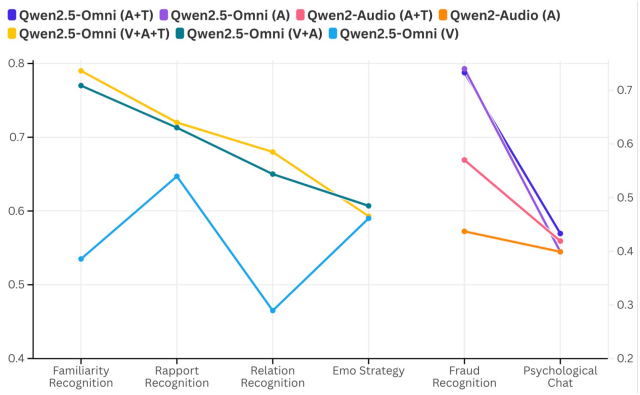

Figure 3: 不同任务中的模态消融。我们可视化了不同模态在从感知到互动各类任务中的贡献。左边的 4 个任务是基于视频的问题，右边的 2 个任务是基于音频的问题。请注意，ASR 转录文本（T）仅用于这个消融研究，未在表格 1 中使用。

与其在视频相关基准测试中的优异表现一致 (?)。值得注意的是，在 L3-L4 高级任务中，所有模型的指标均在 40 到 60 之间变动较小，这表明单凭视觉模态输入不足以为这些任务提供足够信息。

全模型和音频大型语言模型。 音频的加入让全能模型在高级任务（L3，L4）中相较于仅限视觉的大型语言模型具有显著优势，例如关系认知和谎言识别。这种跨模态协同作用在欺诈识别任务中得到进一步强调，其中 Qwen2.5-Omni（0.74）表现优于其专门的音频模型。然而，在更复杂的任务如心理对话中，这种感知优势有所减弱。这强调了当前领先全能模型的一个局限性：主要障碍在于缺乏高级、长背景推理能力，而不是低级感知能力，而这种能力对于真正以人为中心的理解至关重

Models	Avg_{L1} L1	Avg_{L2} L2	FR1	FR2	LD	RR1	RR2	ES	PC
			L3					L4	
Baseline	0.410	0.567	0.770	0.740	0.550	0.713	0.650	0.607	0.399
+ Stg.1	0.555	0.548	0.720	0.557	0.540	0.707	0.620	0.593	0.540
+ Stg.1-2	0.554	0.565	0.775	0.687	0.545	0.693	0.625	0.593	0.625
+ Stg.1-3	0.563	0.603	0.780	0.687	0.555	0.720	0.690	0.620	0.619
+ PE	-	-	0.790	0.763	0.523	0.720	0.625	0.600	0.436

Table 2: 多阶段全模态强化学习和无训练提示增强的评估。我们报告 L1 和 L2 的平均得分，以及每个高级任务的详细得分。PE 代表提示增强。

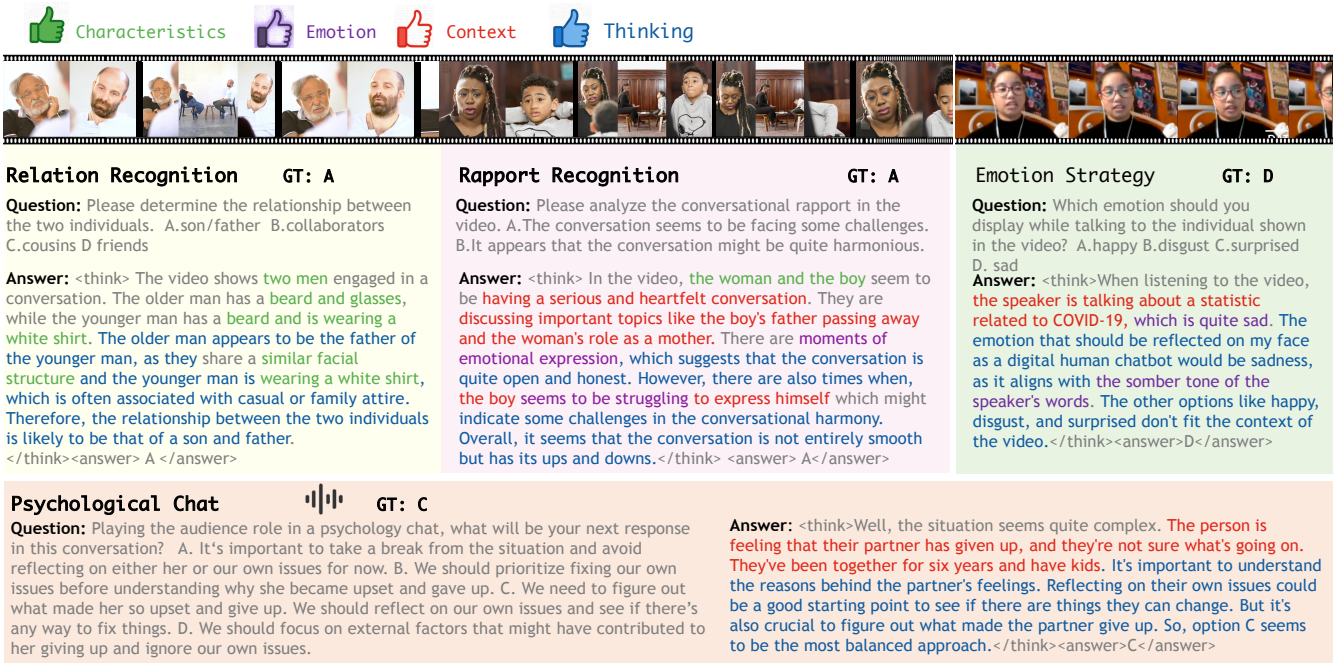

Figure 4: 成功推理的示例。这些案例涵盖四个高层次的感知和交互任务，包括基于视频和基于音频的问题。推理过程都展示了结合特征、情感和上下文的思维，然后提供适当的反馈。

要。

图表 2 中直观展示了全能模型，尤其是 Qwen2.5-Omni，在较高层次任务中显示出明显优于仅视觉 LLM 的优势，凸显了包括听觉在内的多模态感知的关键作用。

为了对不同模态的重要性进行细粒度分析，我们使用 Qwen2.5-Omni 和 Qwen2-Audio 模型进行了系列消融研究。为清晰表述，我们将视频、音频以及视频中的 ASR 转录文本分别记作 V、A 和 T。我们的实验集中在 L3-L4 难度级别中的六个具有挑战性的任务上，其中包括四个基于视频的任务和两个基于音频的任务。我们设计了两个特定的实验设置以探究模态贡献：(1) 在两个模型中通过 ASR 转录文本 (T) 作为附加输入进行增强，(2) 在一个仅可视化设置中评估 Omni 模型，去除音频输入。

图 3 展示了我们的模态消融研究结果。对于视频任务，音频输入 (A) 作为视频 (V) 的一个强大补充，显著提高了诸如融洽关系和关系识别等高层次任务的性能。结合 ASR 文本 (T) 提供的额外好处极小，这表明原始音频已经为这个高级模型提供了大量的语义信息。相比之下，对于基于音频的任务，纯音频模型对文本输入表现出明显的依赖，这突显了其有限的语音理解能力和对明确语义支持的依赖。然而，全模态模型从文本输入中获得的益处有限，这凸显了综合多模态训练的优势。此外，所有模型在与响应相关的任务上的表现都比在感知任务上要差，这强调了在互动场景中提高 MLLMs 响应能力的重要性。

XNUMX 上述评估证实了视觉、听觉和文本信息在高层次任务中都起着重要作用。通过数据观察，我们还发现，适当的沟通反馈依赖于对全模态上下文的全面考虑，以及对对话者需求、情感和个人特征的深入了解。因此，我们认为，基于全模态输入的推理能力是增强 MLLMs 认知和互动能力的关键。在接下来的章节中，我们采用一种多阶段、全模态强化学习的方法来构建一个推理全能模型。

全模态强化学习 我们使用与基准测试无关的数据构建训练集。我们应用了组相对策略优化 (GRPO) (?)，并设计了一种多阶段、全方位的训练方法，使 Qwen2.5-Omni-7B 在可验证奖励的强化学习 (RLVR) 过程中接触到所有模式，增强训练稳定性并逐步加强感知能力。

具体而言，在第一阶段，我们使用纯视频帧和文本问答 (QA) 对进行训练，以建立整合视觉感知的推理能力。

在第二阶段，我们通过基于音频的 QA 进行训练，以发展整合听觉感知的推理能力。最后，我们利用完整的视频音频 QA 来加强结合视觉和听觉感知的推理能力。更多训练细节包含在补充材料中。

如表 2 所示，对于以视觉为中心的 L1 任务，阶段 1 训练已经带来了显著的改进，表明以视觉感知为基础的推理可以增强这类任务的表现。对于与音频相关的任务 PC 和 FR2，阶段 2 训练相比阶段 1 产生了显著的提升，揭示了整合听觉推理的成功之处。大多数任务在完成全部 3 个训练阶段后实现最佳表现。我们抽取了回答正确的例子，发现模型确实能够通过整合特征、情感和上下文信息进行深入思考，如图 4 所示。

我们观察到，通过 RL 训练所引发的成功推理过程展现出一种一致的思维模式：感知关键特征、情感和情境，然后是思考和回应。受此启发，我们相信存在一种无需训练的方法，可以通过提示增强来提高现有 MLLM 的性能。为此，我们设计并测试了以下提示模板，发现它在高级任务中也提升了性能，如表 2 所示。

```
When analyzing the video or audio, focus on
identifying:
   • Key Characteristics: Recognize notable
     features, actions, emotions, or behaviors
     of people.
   • Emotion: Identify the expressed or inferred
     emotional states of individuals
   • Context: Extract relevant dialogue or
     spoken words.
For the following tasks, base your reasoning on
the above elements to draw conclusions.
```

我们引入了 HumanSense 基准，以探索多模态大型语言模型在复杂的人类中心感知和交互场景中的能力。我们提出，全模态推理能够提升多模态大型语言模型在这些任务上的表现。我们的目标是激励社区认识到多模态大型语言模型在推进人工智能交互体验方面的潜力。