人类与大模型对拼接面孔情绪识别的能力差异

赵琳[1,2],李婧婷[1,2]，刘烨[1,2]，马骏驰[1,3]，王甦菁[1,2]

[1]认知科学和心理健康全国重点实验室（中国科学院心理研究所），北京 100101

[2]中国科学院大学心理系，北京 100039

[3]江苏科技大学，镇江 212100

摘　　要 面孔是传递情绪等社会信息的重要载体。人类依赖整体与局部加工等多层次的认知机制高效、准确地识别面部基本情绪；而多模态大语言模型（Multimodal Large Language Models, MLLMs）虽整合了视觉编码组件与语言推理机制，但其加工策略与人类的知觉加工在原理上存在显著差异。将二者的情绪识别能力进行对比有助于理解两者在情绪感知与推理策略上的差异。同时，已有研究提示文本提示词会显著影响 MLLM 的输出，但其在面孔情绪识别情境中的作用仍缺乏系统检验。

基于上述原因，本文旨在探讨面孔情绪识别中的整体与局部特征加工优势，并进一步考察该加工模式在人类与 MLLM 生成的“虚拟参与者”之间是否具有一致性。文章共包含四个实验，结果表明：MLLMs 在识别拼接情绪面孔时表现出区别于人类的局部特征偏好：协调比率低且倾向将图片判断为互斥；提示词的细节化程度与示例图片会显著改变模型的判断倾向与协调比率。综上，研究结果加深了对人类与人工智能在情绪理解路径上的差异的认识，并为人工智能在情绪识别与人机交互领域的应用提供了新的理论参考。

关键词：面部动作单元，互斥性，情绪识别，多模态大语言模型

收稿日期：2025-11-24

基金资助号：62476269, 62276252

通信作者的联系方式：lijt@psych.ac.cn

1 引言

1.1 人类的情绪识别能力

情绪通常被界定为：当个体受到外界刺激时所产生的由生理唤醒、主观体验与行为表达（如面部表情、姿态表情等）共同构成的复杂心理状态(Darwin & Darwin, 1872; Ekman, 1992)。自 Darwin 提出情绪具有进化适应功能以来，大量的跨文化研究发现，人类可以在无语言、语境提示的情况下，仅通过面部表情对基本情绪进行识别，且这种识别具有普适性(Darwin & Darwin, 1872; Ekman & Friesen, 1971; Jack et al., 2014)。因此，面部表情通常被视为人类交流互动过程中具有社会信息的重要载体。大量研究表明，人类能够在短时间内对基本情绪进行相对准确的识别，即使情绪面孔只呈现几百毫秒，观察者仍然能够从面部视觉线索中快速地提取有用情绪线索，并作出情绪类别的判断(Calvo & Nummenmaa, 2016; Martinez et al., 2016; Matt et al., 2021)。情绪的脑电研究进一步揭示了情绪识别过程的快速性，情绪面孔通常在呈现 200 毫秒后即可引发和面部结构编码与区分情绪相关的神经，如 N170 与 EPN，这提示我们情绪识别并不只是完全依赖耗时的高层次推理，也包含快速的知觉加工(Bruchmann et al., 2023; Schindler & Bublatzky, 2020; Weidner et al., 2022)。

人类对情绪面孔的快速识别能力与人类对于面部表情的多层次编码相关，这一多层次过程包含三个关键层次：视觉特征的快速感知、情绪意义的提取与类别判断以及多通道信息整合。首先，面部表情信息在视觉特征提取阶段会被迅速提取与编码。当表情出现时，观察者能够在短时间的情况下捕捉到面部肌肉的变化，进而根据经验作出情绪类别判断(Ziereis & Schacht, 2024)。然而，大量心理学情绪研究指出，人类对于情绪的识别并非仅依赖于对局部特征的线索加工，而是更依赖于对于整张面孔的整体加工（holistic processing），即观察者通过整合各面部部位的空间关系来形成情绪感知(Sun et al., 2023; Tanaka & Farah, 1993)。其次，在情绪意义提取阶段，观察者会根据情绪概念与已有经验，将面部特征与特定的情绪类别（如快乐、悲伤、愤怒等）进行关联(Brooks et al., 2019; Gendron et al., 2014)。最后，观察者会将所观察到的信息与其它线索整合到一起，进一步增加情绪判断的准确性与鲁棒性。这一多层次的加工展现了面部表情识别的复杂性与灵活性，也在一定程度上表明人类对于面部表情的加工并不只是依赖于某一单一信息，而是对于多种感知与认知的整合。

在这种多层次与多认知的背景条件下，当面孔被倒置、分割或组合时，人类情绪识别的准确性以及识别速度都会显著下降(Murphy et al., 2017; Rossion, 2013; Tanaka & Farah, 1993; Yin, 1969; Young et al., 2013)。组合情绪面孔使面部表情信息以一种违背自然生理结构的方式呈现，进而削弱了表情识别过程中的整体加工机制，并且显著增加了识别时的知觉难度。在这种条件下，观察者难以对来自不同面部区域的信息进行较好地整合，进而只能转向依赖局部特征进行

判断(Faustmann et al., 2022; Young et al., 2013)。组合情绪面孔为我们研究多层次的情绪识别加工提供了新的视角。同时，已有研究表明，人类对非自然面部结构配置下的面部表情的高效识别展示出了人类对于多样化情绪信号的敏感性以及高效、精细的识别能力。这种能力既展示出了人类对于面孔情绪信息加工的独特性，又阐明了面部表情这一情绪外显信号在人类社会互动中的重要作用。

虽已有部分研究使用组合情绪面孔来检验表情识别的区域权重，但这些研究多集中于在整脸情况下局部特征的识别，并未系统全面地研究在六种基本情绪的两两组合之下面部某一区域的权重或整合机制。因此，本研究将面孔以上下区域进行重组，形成两类刺激：一类为上下面部来源于相同情绪类别的组合，另一类为上下面部来源于不同情绪类别的组合。需要注意的是，后一类刺激并不是绝对对应于现实生活中的“不协调表情”，其有可能与复合情绪或混合社会表情之间具有一定的相似性。因此，本文中将“同类/异类情绪拼接”视为刺激拼接方式的操作性分类，而将“协调/互斥”定义为参与者基于刺激材料所作出的判断结果。文章旨在通过操纵面部上下区域的情绪信息，探讨不同面孔区域线索的一致性或不一致性对于情绪协调性判断以及情绪识别、情绪识别依据的影响。为此，提出以下假设：

1) 在拼接情绪面孔识别任务中，相较于上下面部来源于不同情绪类别的拼接面孔，人类参与者对上下面部来源于相同的情绪类别的拼接面孔表现出显著高的协调比率；
2) 当呈现面孔为拼接面孔时，基本情绪的识别表现出局部特征依赖，如高兴表情更大程度上依赖于下面部特征；

1.2 大语言模型与人类在面孔情绪识别能力上的差异

已有大量研究证实，人类能够在短时间内对不同的面部表情进行解读，并做出适当地判断。这种识别和判断能力在日常社交中发挥着重要的作用，同时也为我们理解情绪表达和识别背后的机制提供了依据。随着大语言模型（Large Language Models，LLMs）的飞速发展，诞生了一个非常值得探讨的问题：LLMs 是否具有等同于人类的表情识别与理解能力。LLMs 作为人工智能的一部分，它的崛起使得研究者开始关注它在情感识别领域的潜力以及局限性。大语言模型被认为是模拟人类认知功能的载体，它通过大量的数据集进行训练习得数据中的语义关联模式，并在部分任务中表现出类人特征(X. Wang et al., 2023)。尽管当前 LLMs 在语言理解与生成方面均表现出了卓越的能力，但其应用目前只局限于单一的文本模态信息。为了克服其对图像、音频等非语言信息识别的不足，研究者提出了多模态大语言模型（Multimodal Large Language Models, MLLMs)，这种模型可以同时处理文本和视觉等多种模态的信息类型。与传统的 LLMs 相比，MLLMs 不仅可以回答文本问题，还可以基于所呈现的图像信息进行分析与推理，拓展了语言模型的应用范围(Ramesh et al., 2021)。

现在已有部分工作将多模态大语言模型应用于面孔情绪识别任务，例如：在广义情绪识别

任务中对 GPT-4V 进行测试，发现其展示出了强大的视觉理解、整合多模态线索和利用时间信息的能力，这些能力对于情感识别至关重要。但 GPT-4V 主要面向的是通用领域，在识别微表情等需要细致特征的场景中表现不好(Lian et al., 2024)。另外，有研究发现将面部图像的情绪维度（效价-唤醒值、结构化的数值表示）提供给 MLLMs，结果显示 LLMs 难以对基本情绪之外的离散情绪进行分类识别，但是在语义情绪描述任务中表现出了符合人类直觉的较为细腻的文本情感推断表情能力(Mehra et al., 2025)。为全面评估 LLMs 以及 MLLMs 在情绪认知和情绪推理中的表现，研究者们构建了包含有多种情绪识别与推理任务的测试。最新的研究发现即使是在表现最好的模型中，正确情绪识别只占总体的 39.3%，正确思维推理只占总体的 56.0%，这表明当前 MLLMs 的情绪智商仍是不成熟的(F. Zhang et al., 2025)。

为了系统剖析人类与大语言模型在这种复杂情绪加工能力上的本质差异，本研究引入 Marr 的视觉计算三层次理论（Marr's Tri-Level Hypothesis）作为审视框架。Marr 指出，对任何信息处理系统的分析都应涵盖计算理论层、表征与算法层以及物理实现层(Marr, 1982)。由于人类与 MLLMs 在物理实现层（碳基大脑与硅基硬件）截然不同，且在面部情绪识别的计算理论层（即任务目标与逻辑）上具有一致性，因此本文的研究层级明确聚焦于“表征与算法层”（Representation and Algorithm level）。在这一层面上，我们重点探讨二者在面临输入信息时，内部的表征结构以及情绪推断的算法过程有何异同。

认知心理学中，大量经典研究正是通过行为指标（如反应模式、错误分布及混淆关系）来推断个体在信息整合过程中的加工策略。例如，David Navon 提出的 Navon 范式揭示了知觉加工中的整体优先效应(Navon, 1977)，而复合面孔效应（composite face effect）则揭示了面孔加工的整体化特性(Young et al., 2013)，这些结论均主要基于行为数据建立。因此，在明确研究层级的前提下，行为结果可以作为推断宏观信息加工方式的重要依据。本研究延续这一研究传统，采用拼接面孔范式，通过协调比率及情绪判断模式，分析不同类型的参与者在冲突信息情境下的整合方式。

在日常生活中表情是通过多种线索来进行综合理解与识别的，并且表情具有高度多样性。在表征与算法层面上，个体在识别他人的面部表情时，会将视觉线索以及先验知识经验进行整合，使得人类的情绪识别能力既有很高的环境适应性，又能根据现有信息进行相应地推断。以上能力使得人类在面对模糊、冲突、复合的情绪表达时，个体仍能够依托于自身积累的经验表征生成较为有效的情绪推理。而大语言模型与人类的情绪识别存在本质差异。MLLMs 在判断情绪时，并不具备感知和体验过程，它们只是依赖于训练集中已有的数据和图片进行关联统计，对所得到的图片进行模式匹配式的特征表征，所以使得其在面对明确且典型的情绪线索时有较高的识别率，如当输入图片包含愤怒的皱眉时，模型往往能够输出较为准确的结果。然而，一旦所呈现的刺激涉及模糊、冲突或需要基于经验进行识别的表情时，模型往往表现出较差的能

力(Lian et al., 2024; Mehra et al., 2025)。

综上所述，现有研究表明，MLLMs 在情绪识别过程中更多的依据已有训练数据集，选择“最有可能的情绪标签”，而不是对面部表情所传达情绪进行整体的深层分析(Lian et al., 2024; Mehra et al., 2025)。这在一定程度上表明了当前 MLLMs 在处理情绪面部信息时的局限性，尤其是在涉及到模糊表达或冲突表达的任务中，它的表现与人类之间存在显著差异(F. Zhang et al., 2025)。同时，先前有关模糊情绪表达或冲突情绪表达的研究大多使用的为基于合成技术生成的完整面部图像，很少探讨上、下面部区域在 MLLMs 情绪加工中的作用(Bombari et al., 2013; Calder et al., 2000)。基于以上结论，本研究旨在通过对上下面部区域进行拼接来系统地考察 MLLMs 的冲突情绪识别能力，并进一步将其结果与人类参与者进行对比，回答如下假设：

3) 在相同的情绪拼接图片识别任务中，大语言模型的识别准确率显著低于人类参与者的识别准确率；

4) 在对情绪表达拼接图片的识别过程中，大语言模型的情绪判断依据与人类参与者的判断依据部分一致，表明其在表征与算法层的情绪加工策略上具备一定的类人特征；

1.3 提示词对于大语言模型表现的影响

近年来的研究发现提示词的差异会显著影响 MLLMs 的输出结果。提示词（prompt）在信息检索和语言推理任务中发挥关键作用(Kojima et al., 2022; Sahoo et al., 2024)。提示词的设计不仅仅是造成参数上的差异，也会影响模型情绪识别的加工路径和结果一致性。有研究显示，结构化提示（如面部关键点和动作单元）能够显著提高模型对情绪图片的识别准确性(Xu et al., 2023)。在一项 MLLMs 人格与文化差异的研究中发现：当让大模型使用思维链（chain-of- thought，CoT）类型提示词时，模型倾向于给出更加详细的推理过程，并且在任务中呈现的答案具有较高的一致性和可重复性；而在使用直接数字输出（direct-numerical-output）类型提示词时，模型给出的答案更具有随意性或者对于模糊刺激的反应不稳定(Li & Qi, 2025)。此外，有研究表明当提示词较简短时，模型在某些分类任务上给出的答案随意性更高和模糊回答更多(Y. Wang et al., 2023)。然而，有研究指出，在高度主观的任务中，CoT 所得到的结果并不一定比直接回答的结果更好，提示词本身可能并没有真正地改变模型处理任务的方式，而只是帮助模型去进行任务类别的识别(Chochlakis et al., 2025)。

提示词不仅可以决定模型的输出结果，也可以在一定程度上弥补局限。例如，语言提示词可以引导模型聚焦于特定情绪维度以提高模型的分类准确性(Z. Zhang et al., 2024)，而视觉提示（如在图片输入前增加对于特定区域的标注或者高亮处理）则可能进一步的增强模型对于关键面部特征区域的敏感性(Z. Wang et al., 2025; Q. Zhang et al., 2024)。以上这些结果都在一定程度上表明，提示词不仅仅是在言语层面对模型进行控制，更是一个可以直接影响到研究结果的有效性与可靠性的核心变量(Chang et al., 2024)。如果在使用模型进行识别任务时，我们为其提供

多种提示（如语言提示和视觉提示）要求其对所呈现内容进行情绪识别，可能会产生不同的实验结果。

综上所述，本研究引入提示词作为关键变量，考察其对多模态大语言模型在情绪拼接图片识别任务中表现的影响，并据此提出了以下假设：

5) 过多文本提示和示例图片的缺少会影响 MLLMs 对于情绪表达的加工和判断，进而使 MLLMs 的情绪识别能力下降。

2 研究一：人类和多模态大语言模型的情绪识别能力差异与加工策略

本研究旨在通过考察人类与 MLLMs（GPT-4o、Gemma、Qwen）对于面部表情拼接图片的识别准确率差异来探究人类与 MLLMs 在识别面孔情绪时的能力差异。为保证实验结果的稳健性，本研究遵循严格变量控制逐步排除刺激材料本身所带来的干扰，因此，本研究共包含三个实验：研究一 a 首先确定了在标准拼接范式下的人类与 MLLMs 情绪识别的基线差异；研究一 b 通过更改面孔身份，来检验研究一 a 中的结果是否具有跨面孔身份的稳定性；研究一 c 进一步通过去除面孔分割线来考察视觉线索对于研究一 a 中结果的影响。

所有人类参与者提供书面的知情同意书，实验结束后获得一定的报酬。此外，本研究已通过相关伦理委员会的审核，并符合伦理规范。

2.1 研究一 a：人类和多模态大语言模型的情绪识别能力差异

本研究在确立了上下面部区域拼接标准的情况下，考察人类与 MLLMs 在识别拼接情绪面孔时的能力差异。

2.1.1参与者

基于研究目的为探究 MLLMs 与人类的面孔情绪识别能力差异，因此本研究共包含两类参与者：人类参与者和 MLLMs 生成的虚拟参与者。

1. 人类参与者

本研究通过线上对大学生群体进行招募，最终共招募到 41 名参与者（女性 27 名），年龄 18～30 岁（M=22.23 岁，SD=2.40），均为右利手。实验中所有参与者均满足以下条件：视力或矫正视力正常，无精神或神经疾病史，无色盲色弱。所有参与者在充分了解实验目的、实验流程之后，签署知情同意书，并自愿参加实验。且所有参与者在实验结束后均获得相应报酬。

2. 多模态大语言模型生成的虚拟参与者

本研究选取 GPT-4o、Qwen2.5-VL-72B-Instruct（Qwen，https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct）和 Gemma-3-27B-It（Gemma，https://huggingface.co/google/gemma-3-27b-it）作为多模态大语言模型的比较对象。三种模型均属于当前主流通用型 MLLMs，且在模型开放性（闭源/开源）、模型规模及视觉—语言对齐方式

等方面存在差异。选择这三种模型的目的并非对 MLLMs 整体进行穷尽性表征，而是作为具有代表性的案例模型，以考察不同类型 MLLMs 在相同任务中的行为模式是否具有一致性，从而提高研究结论的稳健性。

为了保证实验过程的稳定性与可控性，我们在 MLLMs 中设置了以下参数：请求的超时时间为 60 秒，当模型回复不符合预设规范时允许最多 5 次自动重试。在输出控制方面，为了避免冗余信息并节约计算资源，我们将最大输出长度限制在 150 个 token，同时保持 temperature 参数为 1.0，以确保结果的稳定性更高（该温度在以往实验中已被证实可生成具有较高稳定性的结果）。除上述显式设置的参数外，其余隐式参数均保持默认值。例如，top_p 的参数设定为 1.0，意味着候选词的筛选不进行概率削减。其他未提及的参数同样维持默认状态，以保证实验的可重复性与结果的可比性。在实验过程中，我们严格遵守每个模型的使用方法以及道德伦理，确保大语言模型所生成的内容均为测试任务相关内容。除此之外，本研究充分保护数据隐私，大语言模型生成的所有数据资料均只用于本研究，不向任何相关机构或组织泄漏数据。

3. 虚拟参与者的生成：

在多模态大语言模型条件下，本研究通过重复推理的方式构建“虚拟参与者”。具体而言，对每一张图像，模型均在统一的结构化提示词下完成判断任务，提示词要求模型依次回答五个问题：表情是否协调、协调性等级、情绪极性、情绪类型以及判断所依据的面部区域，并按照固定格式输出结果。为模拟个体间差异，在推理过程中设置 temperature＝1，并启用随机采样且不固定随机种子，使模型在相同输入条件下产生多样化输出。每一次完整推理（覆盖全部 72 张图像）视为一名虚拟参与者，从而在同一模型内部形成多个独立样本（例如，Qwen2.5-VL 共生成 41 轮，对应 41 名虚拟参与者）。需要说明的是，虚拟参与者之间不存在人为设定的提示词差异或角色差异，其变异主要来源于模型生成过程中的随机采样机制，从而避免引入额外的系统性偏差。

本研究得到了通讯作者所在单位伦理委员会的批准(批准编号：H24039)。所有程序遵循《美国心理学会伦理准则》，实验中采集到的数据均进行了匿名化处理，仅研究人员可访问原始数据。

2.1.2实验设计

实验包含三个关键变量：情绪类型、面部表情拼接图片类型和被试类型。其中，情绪类型有 6 个水平：高兴，悲伤，惊讶，恐惧，厌恶，愤怒；面部表情拼接图片有两种类型：同类面部表情情绪表达拼接图（如高兴上面部+高兴下面部），异类面部表情情绪表达拼接图（如悲伤上面部+高兴下面部，具体的情绪组合见表 1）；被试类型有 4 个水平：人类参与者，GPT 模型生成的虚拟参与者，Gemma 模型生成的虚拟参与者，Qwen 模型生成的虚拟参与者。

表 1 研究中不同情绪拼接的组合次数

	高兴	悲伤	惊讶	恐惧	厌恶	愤怒
高兴	6					
悲伤	2	6				
惊讶	3	3	6			
恐惧	2	2	2	6		
厌恶	2	2	2	4	6	
愤怒	3	3	2	2	2	6

2.1.3实验材料

Ekman 等人在面部运动编码系统（Facial Action Coding System， FACS)中所呈现的面部表情动作单元图片是目前流传度最广、认可度较高的图片。为保证我们实验材料所得结果的有效性，实验中的原始刺激材料均来自 FACS 手册中 Ekman 的面部表情图片。在实验中我们共选取了 16 种代表性 AU（由于在实验中我们选取的图片为包含代表性 AU 的面部表情示意图，而非单一的面部运动，所以会包含一些与本研究内容无关的 AU），所选取 AU 及其所归属面部区域如表 2 所示。

表 2 不同情绪的面部特征及判断依据

情绪	上面部	下面部	识别区域
高兴	AU6	AU12，AU28	下面部
悲伤	AU4，AU43	AU15	全脸
惊讶	AU1，AU2，AU5	AU25	上面部
恐惧	AU1，AU4，AU5	AU25	上面部
厌恶	AU7	AU9，AU24	上面部
愤怒		AU16，AU22，AU23	下面部

此外，由于在实验中我们须对面部情绪示意图进行区域划分，我们将所选取的面部情绪示意图总结为以下两种情况：1）面部情绪示意图中所包含的代表性 AU 只存在于一个面部区域。若示意图中只包含 AU4，我们将其命名为 AU_4 示意图，上面部为 AU4，下面部则为 Nan（即无关表情或中性表情）；情绪示意图中包含多种代表性 AU 且上、下面部均存在，那在对上、下面部进行命名时，就用该区域所包含的 AU 表示。如示意图中同时包含 AU6 和 AU12，则上面部用 AU6 表示，下面部用 AU12 表示。

其中，同类面部表情情绪表达拼接图是指将相同情绪类别的上下面部进行拼接所形成的图片，如：高兴上面部+高兴下面部，悲伤上面部+悲伤下面部。异类面部表情情绪表达拼接图是指将不同情绪类别的上下面部进行组合而形成的图片，如：悲伤上面部+高兴下面部，惊讶上面部+高兴下面部。此外，需要注意的是，同类/异类面部表情情绪表达拼接图仅表示刺激材料在构成方式上的分类，而不是假设其对应的是现实生活中的协调/不协调的情绪表达。

2.1.4实验流程

由于本研究中包含两类参与者，且两类参与者间存在较大差异，所以在实验流程方面会存在部分差异。

1. **人类参与者的实验流程**

参与者坐在一个明亮开阔的房间内，眼睛距离显示器 60～80cm，观看电脑显示器（分辨率：2560×1440、刷新率：60HZ）上由 PsychoPy 软件所呈现的面部刺激。每位参与者的任务是对所呈现的面部刺激进行连续五次的判断，整个实验共 360 次按键任务。具体如下，在实验正式开始之前，告知参与者实验流程及被要求的具体操作。每次实验首先在屏幕中央呈现大小为 0.05×0.05 的黑色“+”注视点（500ms），之后面部表情情绪表达拼接图以随机的顺序呈现于屏幕中央。对于每张刺激图片，参与者均需完成 5 个判断任务，刺激图片始终呈现在屏幕上以供参考：

1) 参与者被指示用协调（F 键）或不协调（J 键）之一来标记刺激；
2) 在进行协调性选择后，参与者被指示用五分制从“些许协调/不协调”到“十分协调/不协调”对协调/不协调程度进行评分；
3) 参与者需对拼接图片整体的情绪效价（积极情绪或消极情绪）进行判断，参与者使用键盘按键（F 键：积极情绪，J 键：消极情绪）来对刺激进行反应；
4) 在选择整体情绪类型后，参与者需要从基本情绪中选择出拼接图片所表达出的主要情绪，该任务中被试通过按键来进行反应（1:高兴，2:惊讶，3:生气，4:害怕，5:厌恶，6:伤心）；
5) 最后，参与者需要通过按键来选择他们在进行判断时的依据，如上面部（F 键）或下面部（J 键）。

在参与者作出按键反应之后，呈现下一个判断任务。每张图片的 5 次判断任务全部完成后，再次于屏幕中央呈现黑色注视点“+”（500ms），之后呈现下一张刺激图片，依次循环，直至完成 72 张面部刺激图片共 360 次的按键任务。具体如图 1 所示：

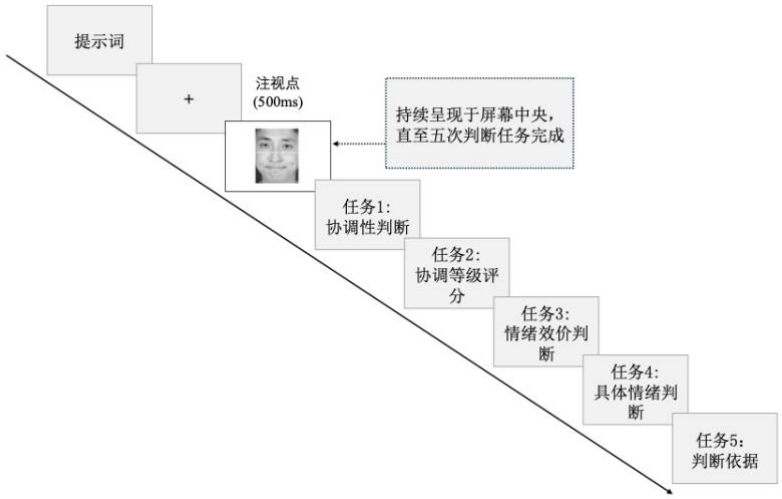


图 1 人类参与者实验流程图

2. 多模态大语言模型的实验流程

对于大语言模型，将提示词(表 3)、示例图片（图 7）和刺激材料发送给大语言模型，要求大语言模型对每张拼接图片进行连续五次的反应。

此外，研究中所设置的情绪效价任务是因为人类在进行情绪加工时会经历两个阶段：粗粒度的效价判断和细粒度的具体情绪判断。所以在研究中，首先让参与者进行效价判断，之后再进行具体情绪的判断，可以在一定程度上保证参与者经历了这个分阶段的认知过程，提升之后具体情绪识别的有效性。由于粗粒度的效价判断和细粒度的具体情绪判断之间存在很高程度的重叠性，所以在后续的数据分析过程中，并未将效价判断纳入分析中。

表 3 提示词

	提示词
问题 1	你认为所呈现的面孔图片中的表情是否可能同时出现在一个人的面孔上，请用是否来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）
问题 2	你认为图片中表情的协调性等级是多少，请用 1-5 之间的数字来进行回答（例如，在上一个选择中你的答案为协调，则 1 代表“些许协调”，2 代表“较为协调”，3 代表“协调”，4 代表“非常协调”，5 代表“十分协调”；不协调如上）
问题 3	你认为图片中的表情属于积极情绪还是消极情绪，请用积极情绪或消极情绪来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）
问题 4	你认为图片中的情绪属于高兴、惊讶、伤心、厌恶、害怕、愤怒中的哪一个？请用以上六个词中的一个词来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）
问题 5	你在判断图片所表达的情绪时，是通过面部哪个区域进行判断的？请用上面部或下面部来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）

2.1.5实验结果

本研究首先对整体数据进行了整理，以考察不同类型的参与者在情绪识别任务中的总体表

现。随后，根据实验设计，依次对人类参与者与三类大语言模型（GPT、Gemma、Qwen）的协调比率、协调性等级评分、情绪混淆、判断依据进行比较分析，检验不同类型的参与者在情绪识别过程中的差异模式。其中，我们将协调比率视为研究的核心因变量，并对其进行了严格的操作性定义：即参与者在协调性选择时将图片选择为协调的比率。值得注意的是，此处的协调比率是指参与者对于刺激的主观判断结果，而并非研究者所定义的刺激属性。计算公式为：协调比率=参与者所给出的协调选择个数/所对应图片类型的总张数（36 张）。如，在同类面部表情情绪表达拼接图（36 张）中，参与者共将 18 张图片判断为协调，则其在该类型图片中的协调比率为 0.5（18/36）;在异类面部表情情绪表达拼接图（36 张）的判断中，若参与者共将 9 张图片判断为协调，则其在该类型图片中的协调比率为 0.25（9/36）。另外，由于 MLLMs 生成的参与者倾向于将所有图片判断为互斥面部表情情绪表达，所以在结果分析中并未对其 AU 组合的协调与互斥情况进行统计分析。

1. 协调比率

在对人类参与者的数据与三类大语言模型各自生成的参与者数据进行对比前，首先要对人类参与者、各自多模态大语言模型生成的参与者数据进行单独的分析。

首先，对四类参与者的协调比率进行描述性统计。通过 **Error! Reference source not found.**，我们得知：四类参与者对于同类面部表情情绪表达拼接图的协调比率均高于异类面部表情情绪表达拼接图。

表 4 参与者判断的协调比率

（其中，1 代表同类情绪表达拼接图，2 代表异类情绪表达拼接图）

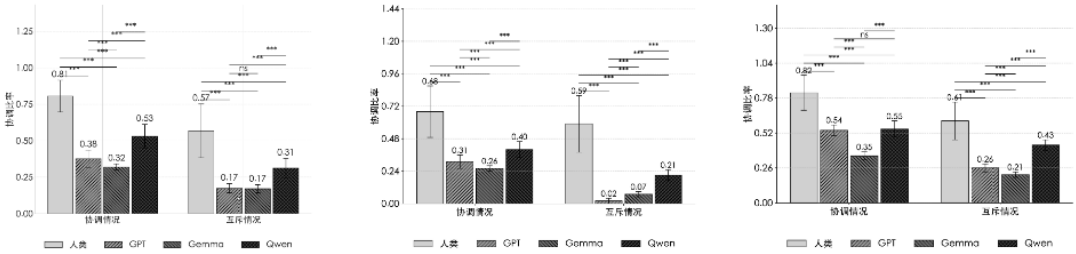
参与者	研究一 a				研究一 b				研究一 c			
	1		2		1		2		1		2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
人类	.81	.11	.57	.18	.68	.19	.59	.21	.82	.13	.61	.14
GPT	.38	.06	.13	.03	.31	.05	.02	.02	.54	.04	.26	.03
Gemma	.32	.02	.13	.03	.26	.02	.07	.02	.35	.03	.21	.02
Qwen	.53	.08	.31	.07	.40	.06	.21	.04	.55	.06	.43	.04

之后，我们对所的协调比率结果进行两因素混合设计方差分析，结果如 **Error! Not a valid bookmark self-reference.**所示。通过方差分析我们发现：图片类型的主效应显著，参与者对于同类面部表情情绪表达拼接图的协调比率显著高于异类面部表情情绪表达拼接图；参与者类型的主效应显著，四类参与者的协调比率从高到低依次为：人类参与者，Qwen 模型生成的虚拟参与者，GPT 模型生成的虚拟参与者，Gemma 模型生成的虚拟参与者；图片类型与参与者类型的交互作用显著。具体而言，简单效应的分析结果如图 2a 所示，即：两种图片的协调比率中均为人类参与者的结果显著高于另外三种 MLLMs 生成的虚拟参与者的结果；Qwen 模型生成的虚拟参

与者的结果显著优于 GPT 模型生成的虚拟参与者和 Gemma 模型生成的虚拟参与者的结果；而 GPT 模型生成的虚拟参与者与 Gemma 模型生成的虚拟参与者两者的对比随面部表情情绪表达拼接图的类型而发生变化，如在同类面部表情情绪表达拼接图的判断中，GPT 模型生成的虚拟参与者的协调比率显著高于 Gemma 模型生成的虚拟参与者，而在异类面部表情情绪表达拼接图的判断中两个模型生成的虚拟参与者的协调比率结果并无显著差异。

表 5 参与者判断协调比率的两因素混合设计方差分析

	变异来源	*df*	*F*	*p*	效应量
研究一 a	图片类型	1	876.24	<.001	.85
	参与者类型	3	7.72	<.001	.84
	图片类型×参与者类型	3	283.68	<.001	.13
研究一 b	图片类型	1	499.19	<.001	.76
	参与者类型	3	248.09	<.001	.82
	图片类型×参与者类型	3	21.19	<.001	.28
研究一 c	图片类型	1	1013.11	<.001	.86
	参与者类型	3	42.68	<.001	.45
	图片类型×参与者类型	3	331.44	<.001	.86



（a）研究一a　　（b）研究一b　　（c）研究一c

图 2 参与者协调比率的多重比较

2. 协调性等级评分

在要求参与者对所呈现的情绪表达拼接图片进行协调性判断后，我们要求其对所呈现图片的协调/互斥性等级进行五点评分。我们对参与者的协调性等级评分进行统计。首先，如图 3 所示，各类参与者具有不同的评分倾向，并且评分的倾向性也会随图片类型而不同，如：在同类面部表情情绪表达拼接图的判断中，人类参与者在评分时更倾向于给出 3–5 分，即认为所判断图片是较高度的协调/互斥；GPT 模型生成的虚拟参与者在评分时更倾向于给出 3 分，即认为所判断图片是中等程度的协调/互斥；Gemma 生成的参与者在评分时更倾向于给出 3 分和 4 分，即认为所判断图片是中等或较强程度的协调/互斥；Qwen 生成的参与者在评分时更倾向于给出 3–5 分，即认为所判断图片要不是中等程度的协调/互斥，要不就是强程度的协调/互斥。在异类

面部表情情绪表达拼接图的判断中，人类参与者的评分倾向与同类面部表情情绪表达拼接图时的倾向保持一致；但 GPT 模型生成的虚拟参与者的评分则转为 2 分和 3 分，即认为所判断图片是轻微或中度的协调/互斥；Gemma 模型生成的虚拟参与者则转为 1 分，即认为所判断图片是些许协调/互斥；Qwen 模型生成的虚拟参与者的评分则转向较为极端，给出的评分多为 1 分和 5 分，即认为所呈现的图片要么是些许协调/互斥，要么就是高度的协调/互斥。

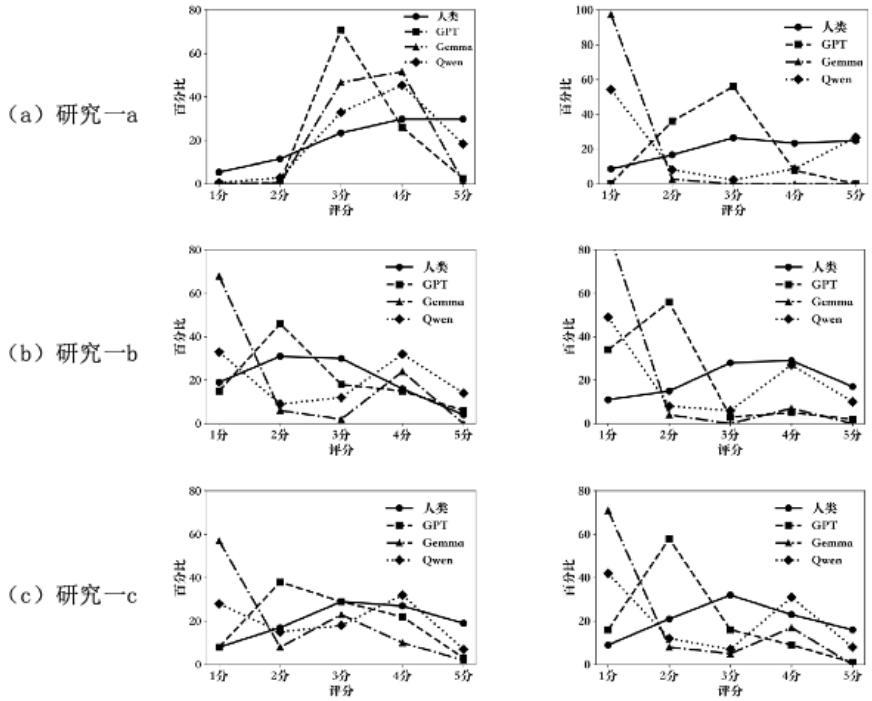


图 3 参与者的分数分布情况

（左图为同类面部表情情绪表达情况下的分数分布，右图为异类面部表情情绪表达情况下的分数分布）

表 6 协调性等级评分的平均分

	参与者	同类–协调	同类–互斥	异类–协调	异类–互斥
研究一 a	人类	3.67	2.87	3.21	3.39
	GPT	3.31	2.84	3.25	2.71
	Gemma	3.53	1.03	3.81	1.03
	Qwen	3.78	2.28	3.93	2.45
研究一 b	人类	2.58	2.55	3.16	3.33
	GPT	3.65	2.09	1.89	1.82
	Gemma	3.92	1.08	3.98	1.05
	Qwen	4.07	2.04	4.07	1.98
研究一 c	人类	3.42	2.81	3.18	3.16
	GPT	3.50	1.82	3.41	1.80
	Gemma	3.39	1.12	3.78	1.10

	Qwen	3.82	1.47	4.03	1.33

之后，我们对参与者的协调性等级评分进行描述统计，平均分如表 6 所示。结果表明：人类参与者整体平均分较高且较为稳定，当其判断正确时会给出更高的分数；对三种多模态大语言模型生成的虚拟参与者而言，当其将图片判断为协调时，其所给出的等级评分的分数会相对较高，而当其将图片判断为互斥时，所给出的等级评分分数相较于协调判断时的分数显著降低。

3. 情绪混淆

对参与者对于同类情绪表达拼接图的情绪选择情况进行统计，发现在情绪选择时存在选择混淆情况。于是，我们对具体的情绪识别选择进行统计，人类参与者结果的混淆矩阵如图 4a 所示（每种情绪各包含 246 次识别）。常见的情绪混淆：29.67%的悲伤被视为厌恶；36.99%的恐惧被误判为悲伤，27.64%的恐惧被误判为惊讶，远超被正确识别为恐惧的数量（17.48%）；41.46%的厌恶被误判为高兴，这一数量远超被正确识别为厌恶的数量（22.36%）；26.42%的愤怒被误判为厌恶。

GPT 模型生成参与者结果的混淆矩阵如图 4b 所示（每种情绪各包含 246 次识别）。GPT 生成参与者的常见情绪混淆：29.27%的悲伤被误判为厌恶；49.19%的恐惧被误判为惊讶，远超被正确判断为恐惧的数量（14.23%）；49.59%的厌恶被误判为高兴，远超被正确判断为厌恶的数量（26.02%）；57.32%的愤怒被误判为厌恶，远超被正确判断为愤怒的数量（17.48%）。

通过图 4c 得知，Gemma 生成参与者的常见情绪混淆：32.52%的悲伤被误判为愤怒；25.61%的惊讶被误判为恐惧；42.28%的恐惧被误判为悲伤，25.20%的恐惧被误判为愤怒，16.26%的恐惧被误判为惊讶，13.01%的恐惧被误判为高兴，远超被正确判断为恐惧的数量（3.25%）；52.03%的厌恶被误判为高兴，33.33%的厌恶被误判为愤怒，13.41%的厌恶被误判为悲伤，远超被正确判断为厌恶的数量（1.22%）；30.49%的愤怒被误判为高兴，23.58%的愤怒被误判为悲伤。

通过图 4d 得知，Qwen 生成参与者的常见情绪混淆：43.09%的悲伤被误判为愤怒；21.54%的惊讶被误判为愤怒，20.33%的惊讶被误判为恐惧；40.65%的恐惧被误判为愤怒，19.92%的恐惧被误判为惊讶，17.48%的恐惧被误判为悲伤，15.04%的恐惧被误判为高兴，远超被正确判断为恐惧的数量（6.50%）；45.12%的厌恶被误判为愤怒，43.09%的厌恶被误判为高兴，7.72%的厌恶被误判为悲伤，远超被正确判断为厌恶的数量（0%）；22.36%的愤怒被误判为高兴。

将人类参与者的结果与三类大模型生成的虚拟参与者的结果进行比较，发现：在 4 类参与者之中厌恶都容易被误判为高兴，恐惧都容易被误判为惊讶；并且，相较于大模型生成的虚拟参与者而言，人类参与者的情绪误判情况更加分散，不会像大模型一样将绝大多数误判为某一种情绪。

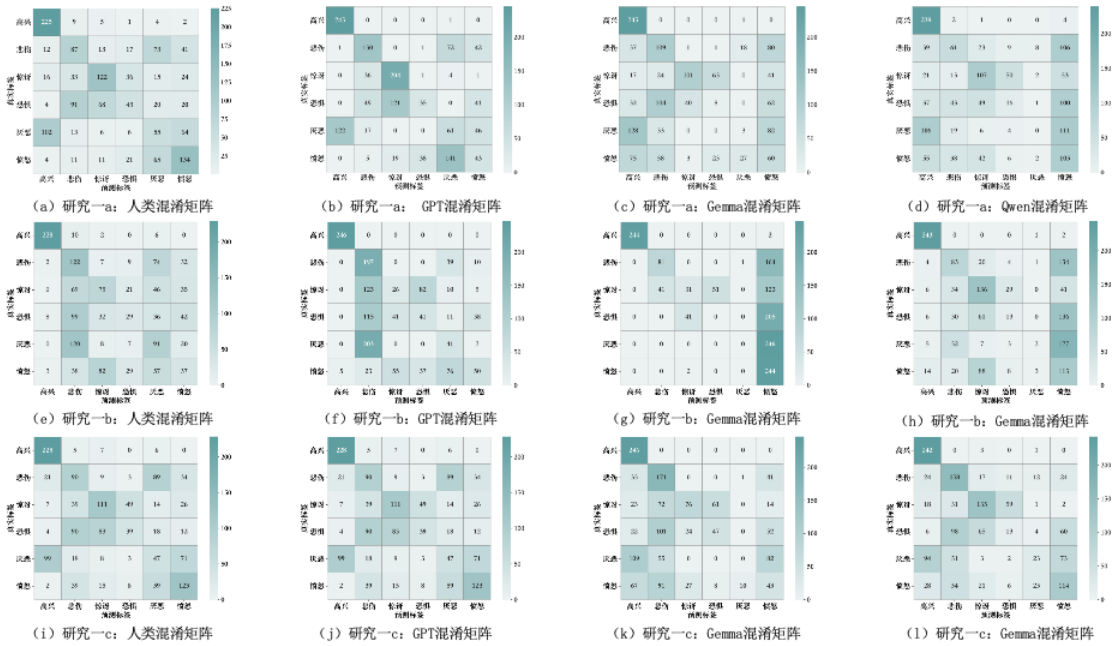


图 4 参与者的情绪判断混淆情况

4. 情绪判断依据

情绪判断会出现情绪混淆状况这一现象使我们产生新的思考：在情绪判断过程中是否存在起决定性因素的面部区域，以及是否每种情绪在判断过程中都包含起决定性因素的面部区域。为更深入的了解情绪判断的依据，我们统计了参与者对于每种情绪的判断依据。首先，我们对所有参与者的判断依据进行描述统计，统计结果如 **Error! Reference source not found.**所示。通过描述性统计，我们得知：对于人类参与者而言，高兴、惊讶、厌恶和愤怒的上、下面部比例之间存在较大差异；在 GPT 模型生成的虚拟参与者所选的判断依据中，所有情绪的上、下面部比例之间均存在较大差异；在 Gemma 模型生成的虚拟参与者所选的判断依据中，高兴和悲伤情绪的上、下面部比例之间存在较大差异；而在 Qwen 模型生成的虚拟参与者的判断依据之中，高兴、悲伤、惊讶和愤怒情绪的上、下面部比例之间均存在较大差异。

之后，我们将情绪判断中每种情绪所选的上面部所占比例作为因变量，对判断依据进行 4*6 的两因素混合设计方差分析，方差分析结果如表 8 所示。通过方差分析，我们得知：情绪类型的主效应显著，这表明不同情绪的判断依据之间存在较大差异；参与者类型的主效应显著，这表明不同参与者在进行情绪判断时的判断依据之间存在较大差异；情绪类型和参与者类型的交互作用显著，这表明不同参与者对于不同情绪进行判断时的判断依据是不同的。

最后，简单效应分析的结果表明：无论是何种参与者，高兴都更容易通过下面部进行识别，惊讶更容易通过上面部进行识别，但其它四种情绪在不同模型中进行判断时的判断依据是不一致的。

表 7 各类参与者在情绪判断时的判断依据

		人类		GPT		Gemma		Qwen	
	情绪	上面部	下面部	上面部	下面部	上面部	下面部	上面部	下面部
研究一 a	高兴	27.67%	72.33%	9.76%	90.24%	0%	100%	8.33%	91.67%
	悲伤	56.41%	43.59%	40.28%	59.72%	0%	100%	69.09%	30.91%
	惊讶	79.55%	20.45%	100%	0%	0%	0%	100%	0%
	恐惧	30.30%	69.70%	100%	0%	0%	0%	100%	0%
	厌恶	48.57%	51.43%	16.67%	83.33%	0%	0%	100%	0%
	愤怒	16.84%	83.16%	0%	0%	0%	0%	33.33%	66.67%
研究一 b	高兴	27.67%	72.33%	9.76%	90.24%	0%	100%	8.33%	91.67%
	悲伤	56.41%	43.59%	40.28%	59.72%	0%	100%	69.09%	30.91%
	惊讶	79.55%	20.45%	100%	0%	0%	0%	100%	0%
	恐惧	30.30%	69.70%	100%	0%	0%	0%	100%	0%
	厌恶	48.57%	51.43%	16.67%	83.33%	0%	0%	100%	0%
	愤怒	16.84%	83.16%	0%	0%	0%	0%	33.33%	66.67%
研究一 c	高兴	27.67%	72.33%	9.76%	90.24%	0%	100%	8.33%	91.67%
	悲伤	56.41%	43.59%	40.28%	59.72%	0%	100%	69.09%	30.91%
	惊讶	79.55%	20.45%	100%	0%	0%	0%	100%	0%
	恐惧	30.30%	69.70%	100%	0%	0%	0%	100%	0%
	厌恶	48.57%	51.43%	16.67%	83.33%	0%	0%	100%	0%
	愤怒	16.84%	83.16%	0%	0%	0%	0%	33.33%	66.67%

表 8 参与者判断依据的混合设计方差分析

	变异来源	*df*	*F*	*p*	效应量
研究一 a	情绪类型	5	57.56	<.001	.34
	参与者类型	3	65.90	<.001	.55
	情绪类型×参与者类型	15	23.40	<.001	.31
研究一 b	情绪类型	5	20.04	<.001	.11
	参与者类型	3	36.46	<.001	.41
	情绪类型×参与者类型	15	39.00	<.001	.42
研究一 c	情绪类型	5	100.89	<.001	.39
	参与者类型	3	7.01	<.001	.61
	情绪类型×参与者类型	15	14.09	<.001	.21

5. AU 组合

在了解了人类参与者的情绪判断依据之后，我们对协调性判断中更加细粒度的 AU 组合进行配对样本 t 检验，为确保差异检验的有效性，对检验结果进行了 Bonferroni 校正，结果发现：

参与者在识别互斥情绪表达拼接图时，更倾向于将 AU4+AU12（悲伤+高兴）、AU43+AU25（悲伤+惊讶）、AU1+AU25（惊讶+恐惧）、AU43+AU24（悲伤+厌恶）、AU1+AU24（厌恶+惊讶）、AU16+AU12（愤怒+高兴）的情绪组合判断为互斥情绪表达组合；在识别协调情绪表达拼接图片时，绝大多数参与者都能够正确的判断为协调，所以并未将协调 AU 组合在该部分列出。

2.1.6MLLMs 注意力热力图分析

注意力热力图生成方法：为进一步探讨模型在面部情绪判断过程中的信息利用方式，本研究对开源模型（Qwen2.5-VL 与 Gemma-3）的注意力权重进行了可视化分析。具体而言，提取模型在生成第一个回答 token 时的注意力矩阵，并定位视觉 token 在输入序列中的位置，将其对应的注意力权重映射回原始图像空间，从而构建面部注意力热力图。为获得稳定的注意力分布模式，对全部 72 张实验图像的注意力矩阵进行累加平均处理，并在归一化前减去全局均值，以突出不同任务条件下的特异性注意力成分。此外，为排除个别图像差异带来的影响，所有图像在统一空间对齐条件下进行处理，以保证不同样本之间的可比性。

注意力热力图结果：注意力热力图分析显示，不同模型在面部信息利用方式上存在显著且稳定的差异（见图 5）。具体而言，Qwen2.5-VL 在各任务条件下均表现出明显的上面部偏向，其注意力主要集中于额头、眉毛及眼睛区域，并形成稳定的顶部高激活模式；相比之下，Gemma-3 呈现出更为分布式的注意力结构，在整个面部范围内均有激活，并在部分任务（尤其是情绪类型判断与判断依据报告）中对嘴部区域表现出显著增强。进一步分析表明，该差异在协调性、情绪强度、情绪极性、情绪类型及判断依据五类任务中均保持一致，未出现跨任务的策略反转。这表明注意力分布模式并非由具体任务需求驱动，而更可能反映模型层面的稳定信息提取策略。

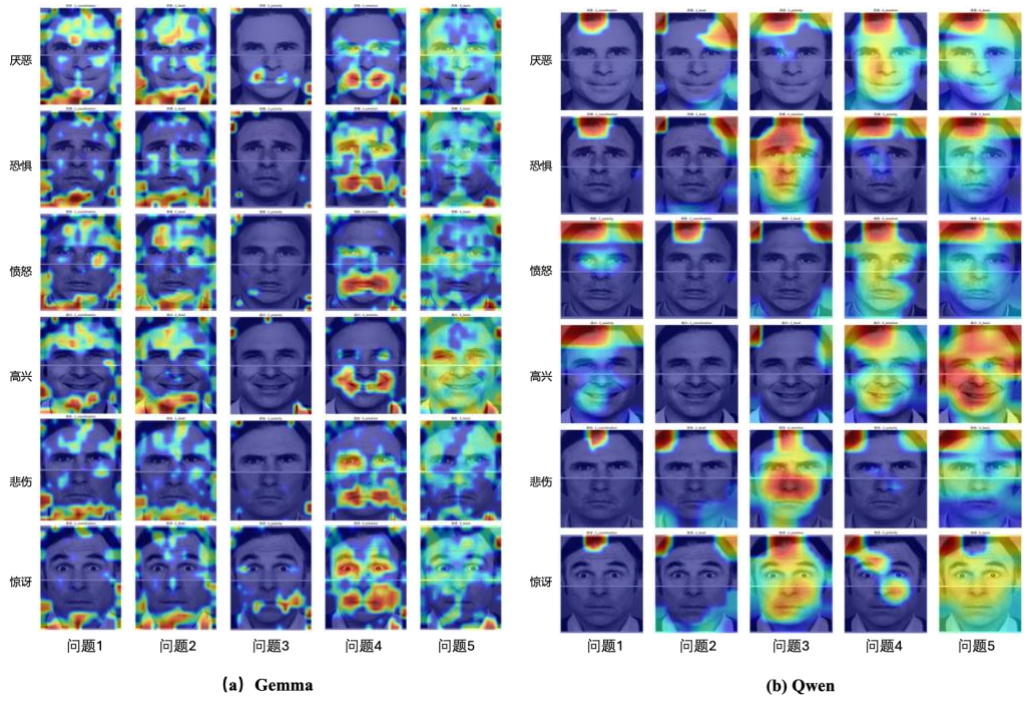


图 5 开源大模型：Gemma 和 Qwen 注意力热力图

2.1.7讨论

在协调分类判断任务中，人类参与者在识别上下面部来源于相同情绪类别的面部表情情绪表达拼接图时的协调比率显著高于在识别上下面部来源于不同情绪类别的面部表情情绪表达拼接图的协调比率。这一结果表明，相较于上下面部来源于不同情绪类别的面部表情情绪表达拼接图，上下面部来源于相同情绪类别的面部表情情绪表达拼接图更容易被整合为统一且连贯的整体。值得关注的是，上下面部来源于不同情绪类别的面部表情情绪表达拼接图并不是意味着其在现实中就是绝对的“不协调”，在某些情境下，这种组合方式可能与复合表情或复杂的社会表情相似。上述结果提示我们，个体在日常生活中接触和学习的展示情绪的面部表情类别具有很高的多样性，因此在面对面部表情拼接图时可能会依赖已有经验进行自动化加工，将上下面部来源于不同情绪类别的面部表情信息进行整理并将其合理化，进而将其误认为协调情绪表达拼接图(Barrett, 2017b; Scherer, 2009)。此外，人类参与者的识别率显著高于三种多模态大语言模型生成虚拟参与者的识别率。这一结果表明：在面部表情，特别是在包含多种面部视觉线索的情绪识别过程中，人类仍然存在显著的优势。此外，由于我们的方差分析是依据协调比率进行的，所以得知：MLLMs 在异类面部表情情绪表达拼接图片的判断识别任务中的表现更好，这可能是由于当前的模型更擅长对情感冲突信号进行检测，而在整合面部信息时的推理能力相对有限。上述结果表明，当前 MLLMs 在复杂的视觉情感任务识别过程中依赖的仍是较低层次的特征匹配，而无法像人类一样对上下文信息进行整合与理解。

在模型之间的比较中，我们发现 GPT 生成的虚拟参与者的识别表现是最好的，但总体协调

比率仍然显著低于人类协调比率。而在模型之中，Gemma 生成的虚拟参与者的表现是最差的。模型之间的差异可能反映了不同模型在情绪语料学习以及情境适应性等方面的局限性。具体而言：GPT-4o 是端到端的的单一模型，视觉和文本的输入均由同一个神将网络进行处理，可能通过在统一的嵌入空间中实现深度的跨模态交互，使模型能够有效利用上下文线索。与之不同的是，Qwen 和 Gemma 则是明确的视觉编码器与语言模型结合的架构，视觉和语言的交互可能受到投影层的限制。并且，Gemma 存在视觉特征固定、令牌压缩、浅层融合等特点，虽然帮助 Gemma 在通用推理的性能提升，但是限制了其在复杂推理任务中的表现。

上述区别表明，即使在行为表现相似的情况下，不同模型也可能通过不同的视觉信息提取路径完成同一任务。这一结果提示，仅依赖行为输出难以充分揭示模型的内部加工路径，不同系统可能通过不同的信息提取方式完成同一任务。因此，在人类与人工系统的比较研究中，有必要引入表征层分析，以避免将行为结果直接等同于加工机制。结合模型架构差异可以推测，不同注意力模式可能源于其视觉编码与注意力机制的结构性差异。例如，全局注意力机制可能更易强化局部关键区域，而混合注意力结构则更倾向于分布式整合信息。

在对图片的协调/互斥性进行等级评分时，我们发现：人类倾向于对自己的判断给出较高的分数，这在一定程度上反映出了人类对于自己判断具有较高的自信度。即使在最初的一致性判断中给出了错误的答案（如将异类面部表情情绪表达拼接图误判为同类面部表情情绪表达拼接图），他们在之后的一致性程度等级评分中仍然给出了较高的分数。这可能是由于部分互斥情绪表达拼接图在整合后所传递的情绪信息与人们在日常生活中所见到的复合社会表情具有高度的相似性，从而在主观体验上被认为是“合理”的表情(Barrett et al., 2011; Russell, 1994)；不同的 MLLMs 生成的虚拟参与者表现出了不同的评分模式：在不同的拼接图片判断任务中，同一个 MLLM 也会表现出不同的评分倾向性。这表明不同的 MLLMs 在处理问题时可能采用的是不同的内部反应策略，这也表现出了当前 MLLMs 训练数据集和泛化能力的差异。而在情绪判断时，我们发现人类和 MLLMs 在一些情绪的判断中表现出了相似的趋势，如厌恶与快乐、恐惧与惊讶之间的混淆，这表明这些情感的面部线索在识别过程中具有一定的混淆性。除了相似的情绪判断混淆性以外，人类的情绪混淆展示出了更为分散的趋势，而 MLLMs 的情绪混淆情况则较为集中且系统化，这说明 MLLMs 对情绪判断的混淆可能是由于其在特征学习过程中将特征与情绪形成了较为强烈的关联。如，模型在学习的过程中，可能将“双眼睁大”这一线索特征与“惊讶”进行了较强的关联，从而导致对恐惧情绪进行了错误分类。

通过对参与者的判断依据进行的分析，我们发现：高兴更容易通过下面部进行识别，而惊讶情绪更容易通过上面部进行识别。这一结果与前人研究中的结果相一致，证明在面部表情的识别过程中不同的情绪确实会存在不同的面部优势区域。这一结果也在一定程度上表明面部表情的识别是一个基于情况进行不同认知资源分配的过程。此外，研究发现：所有类型的参与者

对于高兴情绪的判断均主要依赖于下面部进行识别，这与先前的研究结果相一致。这一发现验证了上述假设：在对情绪表达拼接图片的识别过程中，大语言模型的情绪判断依据与人类参与者的判断依据部分一致，表明其在情绪加工机制上具备一定的类人特征。

2.2 研究一 b、c：面孔身份和分割线对情绪识别的影响

研究一的结果表明，当前 MLLMs 与人类在面孔情绪识别上存在显著差异，但对于造成这一差异的原因尚不明确。在接下来的实验中，我们严格控制变量，逐步验证研究一 a 的结果在不同面孔身份（研究一 1b）与无分割线（研究一 1c）的情况下是否依然成立。

2.2.1参与者

为了保证实验的独立性与对比的严谨性，研究一 b 和研究一 c 在参与者的构成方式上与研究一 a 保持一致，但在具体的采集方式与参数上存在差异。

1. 人类参与者

研究一 b 和研究一 c 均通过线上平台 Pavlovia 独立进行数据收集，两项实验各招募了 41 名大学生参与者。其中，研究一 b 中包含 24 名女性参与者（*M*=20.05，*SD*=3.06）；研究一 c 中包含 24 名女性参与者（*M*=22.05，*SD*=3.45）实验中所有参与者均满足：视力或矫正视力正常，无精神或神经疾病史。所有参与者自愿参与实验，且在实验结束后获得相应报酬。

2. 多模态大语言模型生成的虚拟参与者

为保证实验数据与人类数据的可对比性，每个模型仍均生成 41 名虚拟参与者。本研究中所采用的 MLLMs 以及参数设置除 GPT-4o 以外，均与研究一中保持一致，MLLMs 即 Qwen2.5-VL-72B-Instruct 和 Gemma-3-27B-It，模型设置为超时时间为 60 秒、最多 5 次自动重试、最大输出长度为 150 tokens，以及 temperature 设为 1.0，以及其余参数（如 top_p=1.0）均保持默认设置。因 GPT-4o 版本更新的原因，本研究中采用的 GPT-4o 版本为 gpt-4o-2024-11-20。除此之外，本研究充分保护数据隐私，使用大语言模型生成的所有数据资料均只用于本研究，不向任何相关机构或组织泄漏数据。

本研究得到了通讯作者所在单位伦理委员会的批准(批准编号：H24039)。所有程序遵循《美国心理学会伦理准则》，采集的数据均进行了匿名化处理，仅研究人员可访问原始数据。

2.2.2实验操纵

研究一 b 和研究一 c 的实验设计、实验流程与研究一 a 保持一致。实验之间的区别为刺激材料类型的不同。

1. 研究一 b：面孔身份对人类与 MLLMs 情绪识别能力差异的影响

在实验中，我们采用了不同于研究一 a 的面孔身份刺激材料。由于目前的真实面孔表情数据库中的素材图片并不能够确保其面部动作单元的标准化，很难确保拼接后的素材具有与 Ekman 同等的标准化程度，所以为保证不增加额外变量，本研究使用 Gemini 基于研究一的面孔

身份刺激材料生成新的面孔图像。

具体而言，我们以研究一中的面孔为参照，通过提示词约束 Gemini 的图片生成过程，使新生成的面孔身份材料在一定程度上与原有刺激材料保持一致：1）情绪面孔图片的 AU 组合保持一致；2）上下面部的分割方式与拼接位置保持一致；3）新生成图片的整体呈现格式保持一致。在这些基础上，只对面孔身份这一变量进行变更，以确保与研究一结果的可对比性。此外，在本研究中，我们只引入了一位新的面孔身份，而没有使用多个不同的面孔身份。原因包含以下两点：首先，本研究的目的是想要检验研究一中的实验结果是否受到特定面孔身份的影响，因此只引入一个新的面孔身份可以在最大程度控制额外变量的前提下来检验面孔身份对实验结果模式的影响；其次，如果同时增加了多个新的面孔身份，可能会产生一些额外的差异，如五官比例、面部纹理等，从而使得实验之间的可对比性降低，以及不能够清晰地检验面孔身份本身所带来的影响。因此，本研究中只引入了一个新的面孔身份，以确保实验结果的内部效度。

在生成面孔身份情绪刺激图片时的提示词以英文方式输入，以下为对应的提示词中文翻译示例：一张年龄相近的白人男性的逼真、高分辨率正面肖像照。他呈现出一种非常特定的面部表情：眉毛下垂且紧蹙（眉下压肌），上眼睑大幅上提，露出虹膜上方的巩膜（上睑提肌），下唇向下拉伸，清晰露出下排牙齿（下唇下压肌）。纯净中性背景，平光棚内布光，正面对镜头拍摄。宽高比 3:4。此外，照片中的人需要与我所提供的示例图上的除面部表情、面孔身份以外的信息以及背景信息保持一致。

总体而言，研究中所使用的实验材料在严格匹配研究一中刺激材料的情绪表达组合和拼接方式的前提下，仅改变面孔身份，确保能够直接检验面孔身份对参与者情绪识别能力差异的影响。实验中所采用的情绪面孔如图 6 左侧图片所示。

2. 研究一 c：分割线对人类与 MLLMs 情绪识别能力差异的影响

研究一 c 中刺激材料是基于研究一 a 所使用的面孔拼接图片进一步处理所得到的。在研究一 a 和研究一 b 中，上下面部之间具有一条明显的分割线，以区分不同的面部表情区域。在研究一 c 中，为了尽可能降低视觉线索所带来的影响，在保持 AU 组合方式、上下面部来源以及图片风格等因素不变的前提下，使用大模型的图片处理功能对刺激图片进行处理，去除上下面部之间的分割线，使两部分在视觉上形成更加自然的面孔结构。最终得到的刺激图片数量与研究一保持一致，并且同样包含同类情绪拼接图片和异类情绪拼接图片两种类型。实验中所采用的情绪面孔如图 6 右侧所示。

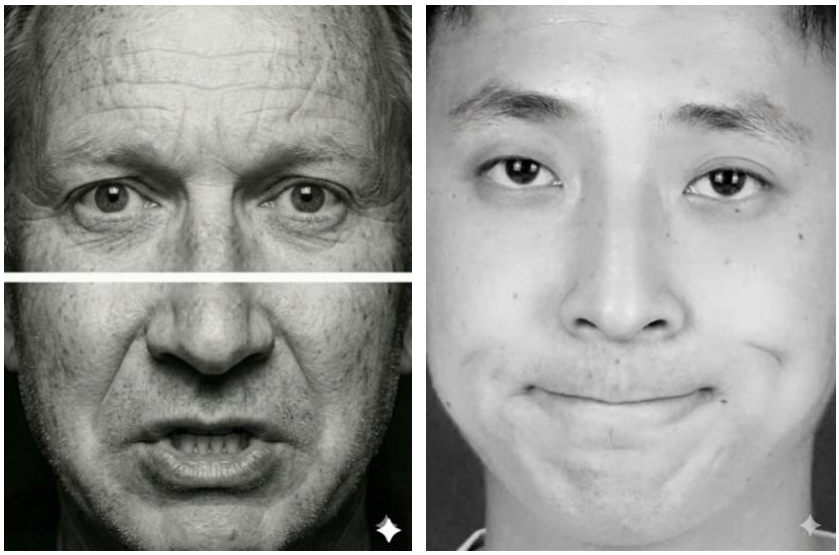


图 6 研究二面孔情绪图片示意图

（左图为研究一 b 中所用素材示例图片，右图为研究一 c 中所用素材示例图片）

2.2.3实验结果

所有数据在统计分析前均经过整理与编码，以确保格式一致并便于后续分析。本研究采用 Python 对数据进行预处理与统计分析。研究一 b 和研究一 c 分别对实验材料进行了不同的操纵，但实验结果与研究一 a 具有较高的一致性。

1. 协调比率

通过表 4、表 5 与图 2b，我们可以得知：与研究一 a 相一致，图片类型与参与者类型的主效应以及交互作用均显著；人类的协调比率仍显著高于 MLLMs 生成的虚拟参与者。四类参与者对于同类面部表情情绪表达拼接图的协调比率均高于异类面部表情情绪表达拼接图。此外，在研究一 c 中所有参与者的协调比率都出现了一定程度的变化，尤其是 MLLMs 的协调比率得到了提高。

2. 协调等级评分

在等级评分任务中，MLLMs 与人类表现出完全不同的评分策略，并且这种差异并未因刺激材料的更改而发生变化。通过图 3 得知，在研究一 b 和研究一 c 中，人类参与者整体评分大多集中于 3–5 分且较为稳定，但当其判断正确时，其会给出更高的分数。与此同时，MLLMs 的评分策略存在较大内部差异：Gemma 生成的虚拟参与者在所有的评分中都更倾向于给出 1 分，即认为所判断图片是轻微程度的互斥/协调；GPT 生成的虚拟参与者更倾向于给出 1 分或 2 分，即认为所判断图片是较低程度的互斥/协调；Qwen 生成的虚拟参与者评分出现较强的两极分化，给出的评分多集中于 1 分或 4 分，即认为所判断的图片是轻微/较高强度互斥/协调。此外，通过表 6 得知，当其将图片判断为互斥时，其评分的平均分会出现较大程度下降。

3. 情绪混淆矩阵

通过对具体情绪识别情况的统计情况（图 4e 至图 4l），我们发现，与研究一 a 相一致，人类与 MLLMs 在情绪误判的分布上呈现出“分散性”与“集中性”的鲜明对比。

人类的情绪混淆较为分散，如在研究一 b 中，人类将 30.08%的悲伤被视为厌恶；28.05%的惊讶被视为悲伤；33.33%的愤怒被误判为惊讶，23.17%的愤怒被误判为厌恶，远超被正确判断为愤怒的数量（15.04%）等，误判较为分散，并不完全局限于某一情绪。而 MLLMs 的误判情况则较为集中，存在将不同情绪系统地误判为某一特定情绪的现象。如在研究一 b 中，Gemma 模型生成的虚拟参与者将 100%的厌恶以及 83.33%的恐惧全部误判为愤怒，GPT 生成的虚拟参与者将 82.52%的厌恶误判为悲伤；在研究一 c 中，GPT 生成的虚拟参与者将 50%的厌恶误判为高兴，50.41%的恐惧误判为惊讶。

4. 情绪判断依据

对参与者的情绪判断依据进行统计，统计结果如表 7 所示。通过描述性统计可以发现，四种类型的参与者在两种刺激条件下都更多地通过下面部进行高兴情绪的识别。在悲伤、惊讶等情绪中，MLLMs 与人类所依据的判断依据具有较大差异，并且存在不稳定的情况。如，在研究一 b 中，人类和 GPT 生成的虚拟参与者均主要通过下面部进行悲伤情绪的识别，Qwen 生成的虚拟参与者主要通过上面部对悲伤情绪进行识别；而在研究一 c 中，人类则转为和 Qwen 生成的虚拟参与者一样，通过上面部进行悲伤情绪的识别。

之后，对判断依据进行方差分析，分析结果如表 8 所示。方差分析结果与研究一 a 保持一致，即情绪类型主效应显著、参与者类型主效应显著、情绪类型与参与者类型之间的交互作用显著。

2.2.4讨论

研究一 b 通过更换刺激材料中的面孔身份，研究一 c 通过去除上下面部之间的分割线，进一步对研究一 a 中观察到的 MLLMs 与人类在情绪识别能力上存在的差异进行检验。总体结果表明，在新的面孔身份以及去除分割线的条件下，研究一 a 中的结果得到了检验。具体来讲，无论是人类参与者还是三类由 MLLMs 生成的虚拟参与者，均表现出了在判断同类面部表情情绪表达拼接图时的协调比率显著高于异类面部表情情绪表达拼接图。此外，人类参与者的协调比率显著高于三类 MLLMs 生成的虚拟参与者，模型之间的顺序也相对一致，即 Qwen 优于 GPT 和 Gemma。以上结果表明 MLLMs 与人类在情绪识别任务中的差异并不是由特定的面孔身份和分割线所提供的视觉线索导致的，这种差异具有跨条件的稳定性。

与研究一 a 相比，研究一 b 中人类参与者的协调比率轻微下降，并且图片类型之间的差异减小。而 MLLMs 的变化较小，这在一定程度上表明，MLLMs 在进行情绪判断时更多的是依据从数据集中所学习到的特征进行局部判断，而较少利用整体信息进行识别。同时，在协调等级评分与情绪混淆模式中，MLLMs 依旧表现出明显的评分差异和将多种情绪集中误判为某一情

绪类别的集中性误判趋势，而人类的评分则较为稳定并且情绪误判较为分散。这一结果进一步说明，MLLMs 在进行情绪判断时可能更多地依赖于局部特征进行分类，一旦所需判断的图片中包含该特征，模型便会将其归为某一固定类型。

在研究一 c 去掉面部分割线之后，所有参与者的协调比率都出现了一定程度的变化，尤其是 MLLMs 生成的虚拟参与者的协调比率得到了明显的提高。但人类参与者在整体表现上仍显著优于三类 MLLMs 生成的虚拟参与者，模型之间的顺序也基本一致。上述结果表明，分割线强化了上下面部之间的分割线，使得模型将上下面部看作是独立的信息来源，进而将其判断为互斥。当分割线被移除之后，刺激材料的连贯性更高，进而提高了协调判断的比例。虽然模型的整体表现发生了变化，但是 MLLMs 与人类之间的差异仍然是与研究一相一致的，这在一定程度上表明，分割线的存在并不是导致参与者之间存在差异的主要因素。

综合来看，上述两个实验证明面孔身份与面部分割线并不是导致 MLLMs 与人类在情绪识别能力上存在差异的主要因素。

2.3 讨论

本研究通过三个递进的实验：识别能力差异检验、控制面孔身份、去除分割线，系统地探讨了 MLLMs 与人类在面部情绪拼接图片识别能力上存在的差异以及潜在的加工策略。总体来讲，三个实验的结果具有一致性，均表明：在包含线索冲突或复杂视觉线索的情绪识别任务中，人类能够对这些信息进行整合，而当前的 MLLMs 仍主要依赖于局部特征进行加工，并且这种 MLLMs 与人类之间的差异具有跨面孔身份和分割线条件的稳定性。

在识别面孔情绪表达拼接图片时，人类的协调比率显著高于 MLLMs 的协调比率。这表明人类在加工观看到的面孔图片时，采用的是整体加工策略，并且能够结合自身已有知识经验，将所看到的面孔合理化并整合为一个整体。相比较之下，MLLMs 则更多地依赖于从大量数据集中所习得特征进行局部特征匹配。研究一 b 和研究一 c 的结果进一步验证了 MLLMs 与人类在进行情绪加工时的策略差异。当面孔身份发生变化时，人类参与者的协调比率轻微下降，而依赖局部特征进行判断的 MLLMs 则表现相对稳定；当上下面部之间的分割线被去除时，由于面孔的完整性得到了提高，MLLMs 的协调比率有轻微上升，这说明分割线会在一定程度上影响 MLLMs 的判断，但即使排除了这一因素，MLLMs 与人类在情绪识别能力上的差异仍然存在。这充分证明，MLLMs 与人类在情绪识别能力上的差异并不是由特定的面孔身份或分割线这一视觉线索所导致的，而是所采取的加工策略不同。

在协调性等级评分中，人类倾向于对自己的判断给出较高的评分，展现出了其对自己的判断较为自信；而 MLLMs 的评分则较为刻板并且易极端化，对自己的判断自信程度较低。在具体的情绪判断中，人类的情绪误判均表现出较为分散的趋势，而 MLLMs 的情绪误判则表现出高度的集中化。虽然所有类型的参与者都更易通过下面部进行高兴情绪的识别，但在判断其他

情绪时，MLLMs 则表现出了将某一局部特征与特定情绪进行强相关的倾向。如，在很大程度上将恐惧误判为惊讶，将厌恶误判为高兴等。以上现象表明，MLLMs 在进行情绪判断时更多地是依赖于面部的某一线索，忽视了上下面部整体的信息进而导致集中式误判。

此外，不同 MLLMs 之间的表现也存在差异，这进一步揭示了模型内部表征的策略多样性。为进一步分析不同多模态大语言模型在面部情绪加工中的潜在差异，对所选模型的视觉编码与注意力机制进行了结构层面的梳理，结果表明：不同多模态大语言模型在面部情绪识别任务中可能采用不同的信息整合策略。具体而言，Qwen2.5-VL-72B-Instruct 采用 NaViT（Native Vision Transformer）视觉编码器，支持动态分辨率输入，不同尺寸图像将被划分为数量不等的视觉 token，并以连续区间形式嵌入到文本序列中。其语言主干采用全局自注意力机制，使任意视觉 token 均可直接参与输出 token 的计算，从而在结构上具备对特定局部区域进行选择性强化的能力。Gemma-3-27B-It 则采用 SigLIP 视觉编码器，将输入图像统一压缩为固定数量的视觉 token。其语言主干采用局部滑动窗口注意力与全局注意力交替排列的混合结构，使视觉信息需经过逐层整合后才能影响最终输出，这一机制在结构上更倾向于分布式的信息整合。GPT-4o 作为闭源模型，其具体结构与中间表征不可获取，仅能通过 API 接口获得输出结果，因此无法进行同等粒度的内部表征分析。上述差异表明，不同模型在视觉信息进入决策路径的方式上存在本质不同，这为后续从表征层分析其信息利用方式提供了结构性依据。

综上所述，本研究中所包含的三个实验证明：当前 MLLMs 在面部情绪识别中缺乏类似于人类的整体加工能力，并且这种局限性无法通过单纯的改变面孔身份或去除分割线来克服。但已有研究表明，MLLMs 的判断可能会受到任务提示的影响。不同的提示词可能导致其对于任务的理解方式不同，进而出现不同的结果。因此，在排除了面孔身份与视觉线索这些因素之后，我们想要探讨模型的情绪识别是否会受到提示词的影响。

3 研究二：提示词对于多模态大语言模型情绪识别性能的影响

上述三个研究结果表明，当前多模态大语言模型在面部情绪识别任务中的表现与人类仍存在显著差异。此外，已有研究提示，大语言模型的输出可能受到提示词表述方式的影响。然而，针对面部情绪识别领域，提示词是否会影响模型的识别准确性仍缺乏系统性证据。基于此，本研究将提示词作为关键自变量，进一步考察提示词对模型面孔情绪识别性能的影响。

3.1 参与者

研究二的主题为提示词对于多模态大语言模型情绪识别性能的影响，所以在研究二中，我们的参与者均为多模态大语言模型生成的虚拟参与者。其中，模型的选择与研究一 a 保持一致，均为 GPT、Gemma、Qwen。不同的是，在研究四中，每种大语言模型都需要生成 164 位虚拟参与者。每种 MLLMs 生成的 146 位虚拟参与者被随机分为 4 组（每组 41 人），之后 4 组虚拟参

与者被随机分至不同的指导语条件下，根据所分配指导语完成情绪识别任务。

3.2 实验设计

研究二在研究一 a 变量的基础上新增一个关键变量-提示词类型。提示词类型共包含四种维度：提示词 1，提示词 2，提示词 3，提示词 4。由于所选取的模型均具有图片识别能力，所以提示词中包含纯文本提示词和文本与示例图片相结合的提示词。其中，提示词 1 和提示词 2 为纯文本提示词，提示词 3 是在提示词 1 的基础上增加了示例图片，提示词 4 是在提示词 2 的基础上增加示例图片。提示词文本具体内容以及示例图片如表 9 和图 7 所示。

表 9 两种文本提示词

提示词 1	提示词 2
你认为所呈现的面孔图片中的表情是否可能同时出现在一个人的面孔上，请用是否来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）	请忽略上面部与下面部之间的空白分割线以及图片的灰度分布不均匀和拼图问题等情况，将上、下面部结合起来视为完整的人脸，回答接下来的问题：你认为所呈现的面孔图片中的表情是否可能同时出现在一个人的面孔上，请用是否来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）
你认为图片中表情的协调性等级是多少，请用 1－5 之间的数字来进行回答（例如，在上一个选择中你的答案为协调，则 1 代表“些许协调”，2 代表“较为协调”，3 代表“协调”，4 代表“非常协调”，5 代表“十分协调”；不协调如上）	请忽略上面部与下面部之间的空白分割线以及图片的灰度分布不均匀和拼图问题等情况，将上、下面部结合起来视为完整的人脸，回答接下来的问题：你认为图片中表情的协调性等级是多少，请用 1－5 之间的数字来进行回答（例如，在上一个选择中你的答案为协调，则 1 代表“些许协调”，2 代表“较为协调”，3 代表“协调”，4 代表“非常协调”，5 代表“十分协调”；不协调如上）
你认为图片中的表情属于积极情绪还是消极情绪，请用积极情绪或消极情绪来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）	请忽略上面部与下面部之间的空白分割线以及图片的灰度分布不均匀和拼图问题等情况，将上、下面部结合起来视为完整的人脸，回答接下来的问题：你认为图片中的表情属于积极情绪还是消极情绪，请用积极情绪或消极情绪来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）
你认为图片中的情绪属于高兴、惊讶、伤心、厌恶、害怕、愤怒中的哪一个？请用以上六个词中的一个词来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）	请忽略上面部与下面部之间的空白分割线以及图片的灰度分布不均匀和拼图问题等情况，将上、下面部结合起来视为完整的人脸，回答接下来的问题：你认为图片中的情绪属于高兴、惊讶、伤心、厌恶、害怕、愤怒中的哪一个？请用以上六个词中的一个词来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）
你在判断图片所表达的情绪时，是通过面部哪个区域进行判断的？请用上面部或下面部来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）	请忽略上面部与下面部之间的空白分割线以及图片的灰度分布不均匀和拼图问题等情况，将上、下面部结合起来视为完整的人脸，回答接下来的问题：你在判断图片所表达的情绪时，是通过面部哪个区域进行判断的？请用上面部或下面部来进行回答（如无法进行判断，请说出一个你认为可能性较大的答案）

图 7 示意图

3.3 实验流程

研究二的实验流程在总体上延续了研究一 a 中大语言模型部分的设计，但在部分流程上进行了调整。具体来讲，研究二要求大语言模型分别生成四组参与者，每组包含 41 名参与者。这四组参与者被随机分配至不同的指导语条件，并在相应条件下完成实验任务。将指导语类型设置为被试间变量可以确保我们能够更加准确地检验出提示词对于实验结果的影响。

3.4 结果

本研究的参与者包括 164 次 GPT 模型生成的虚拟参与者参与者数据、164 次 Gemma 模型生成的虚拟参与者数据以及 164 次 Qwen 模型生成的虚拟参与者数据。每位参与者需对 72 张照片分别进行五次连续回答，共计 360 次回答。所有数据在统计分析前均经过整理与编码，以确保格式一致并便于后续检验。本研究采用 Python 对数据进行预处理与统计分析。

3.4.1协调比率

在将 MLLMs 的不同提示词进行对比之前，我们首先介绍一下各 MLLMs 生成的参与者的各自的结果。通过表 10，我们可以得知对于三种 MLLMs 来讲，它们对于协调情绪表达拼接图的协调比率均高于互斥情绪表达拼接图。

表 10 MLLMs 在各提示词条件下协调比率

（其中，1 代表同类情绪表达拼接图，2 代表异类情绪表达拼接图）

参与者	提示词 1				提示词 2				提示词 3				提示词 4			
	1		2		1		2		1		2		1		2	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
GPT	.08	.03	.01	.02	.25	.05	.10	.04	.38	.06	.17	.03	.31	.09	.15	.06
Gemma	.09	.02	.11	.01	.06	.02	.09	.02	.32	.02	.17	.03	.09	.03	.11	.04
Qwen	.24	.05	.10	.03	.13	.05	.08	.04	.53	.08	.31	.07	.42	.05	.30	.05

之后，我们对实验所得协调比率进行三因素混合设计方差分析，结果如表 11 所示。方差分析的结果表明：参与者类型的主效应显著，即参与者的协调比率之间存在显著差异，协调比率从高到低依次为 Qwen 模型生成的虚拟参与者的协调比率、GPT 模型生成的虚拟参与者的协调比率、Gemma 模型生成的虚拟参与者的协调比率；提示词类型的主效应显著，即提示词类型会对参与者的协调比率产生影响；图片类型的主效应显著，同类面部表情情绪表达拼接图的协调比率显著高于异类面部表情情绪表达拼接图；此外，两因素交互作用和三因素交互作用均显著。

表 11 MLLMs 协调比率的三因素混合设计方差分析

变异来源	*df*	*F*	*p*	效应量
参与者类型	2	488.64	<.001	.89
提示词类型	3	946.83	<.001	.89
图片类型	1	1224.69	<.001	.91
参与者类型×提示词类型	6	150.85	<.001	.72
参与者类型×图片类型	2	195.09	<.001	.77
图片类型×提示词类型	3	165.26	<.001	.58
参与者类型×提示词类型×图片类型	6	26.39	<.001	.31

之后，对交互作用进行简单效应分析，分析结果如图 8 所示：对于三种 MLLMs 来讲，均为提示词 3（提示词 1+示例图片）的协调比率显著高于其他三种提示词，之后为提示词 2 的协调比率高于另外两种类型的提示词，而另外两种提示词类型在不同的模型间表现出了不同的结果，如在 GPT 模型生成的虚拟参与者中为提示词 2 的协调比率大于提示词 1；而在 Gemma 模型与 Qwen 模型中则为提示词 2 的协调比率小于提示词 1。

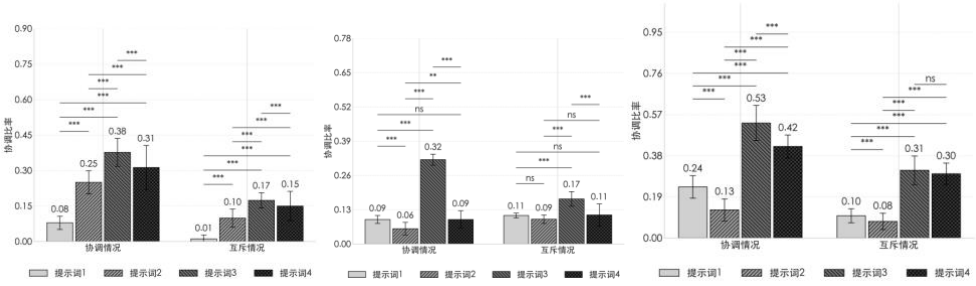


图 8 MLLMs 生成的参与者的协调比率多重比较

（从左到右依次为：GPT 模型内部提示词类型间比较，Gemma 模型内部提示词类型间比较，Qwen 模型内部提示词类型间比较）

3.4.2协调性等级评分

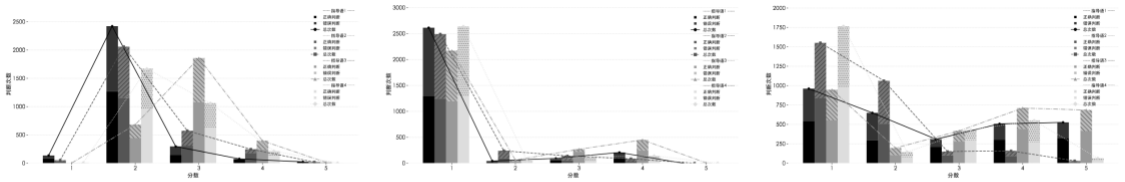
在分析了三种大语言模型在四种提示词情况下的协调比率之后，我们对其协调性等级评分的情况分别进行了统计。接下来依次为 GPT 模型在四种提示词情况下分别生成的参与者的分数分布情况、Gemma 模型在四种提示词情况下分别生成的参与者的分数分布情况、QWen 模型在四种提示词情况下分别生成的参与者的分数分布情况。

首先，我们对 GPT 模型在四种提示词情况下分别生成的参与者的分数分布情况进行统计。通过图 9a 的对比，我们可以得知：无论在哪种提示词情况下，GPT 生成的分数都更集中于 2 分和 3 分。

之后，我们对 Gemma 模型在四种提示词情况下分别生成的参与者的分数分布情况进行统计。通过图 9b 对比得知：无论在何种提示词情况下，Gemma 都更倾向于给出 1 分的评分。

最后，我们对 Qwen 模型在四种提示词情况下分别生成的参与者的分数分布情况进行统计。通过图 9c 对比得知： Qwen 模型生成的参与者等级评分较为分散，并且除了较集中于 1 分之

外，其余的分数出现情况较为复杂，并未像其它大语言模型一样固定于某一分数。



(a) GPT 模型生成的参与者间的一致性等级评分分数分布　(b) Gemma 模型生成的参与者间的一致性等级评分分数分布　(c) Qwen 模型生成的参与者间的一致性等级评分分数分布

图 9 大语言模模型生成的参与者间的一致性等级评分分数分布

综上所述，GPT 和 Gemma 两种大语言模型生成的参与者的评分趋势是一致的，无论是何种提示词；相较于其它大语言模型固定于某一分数，Qwen 这一大语言模型生成的参与者评分除集中于 1 分之外，无固定的评分趋势。

3.4.3情绪判断

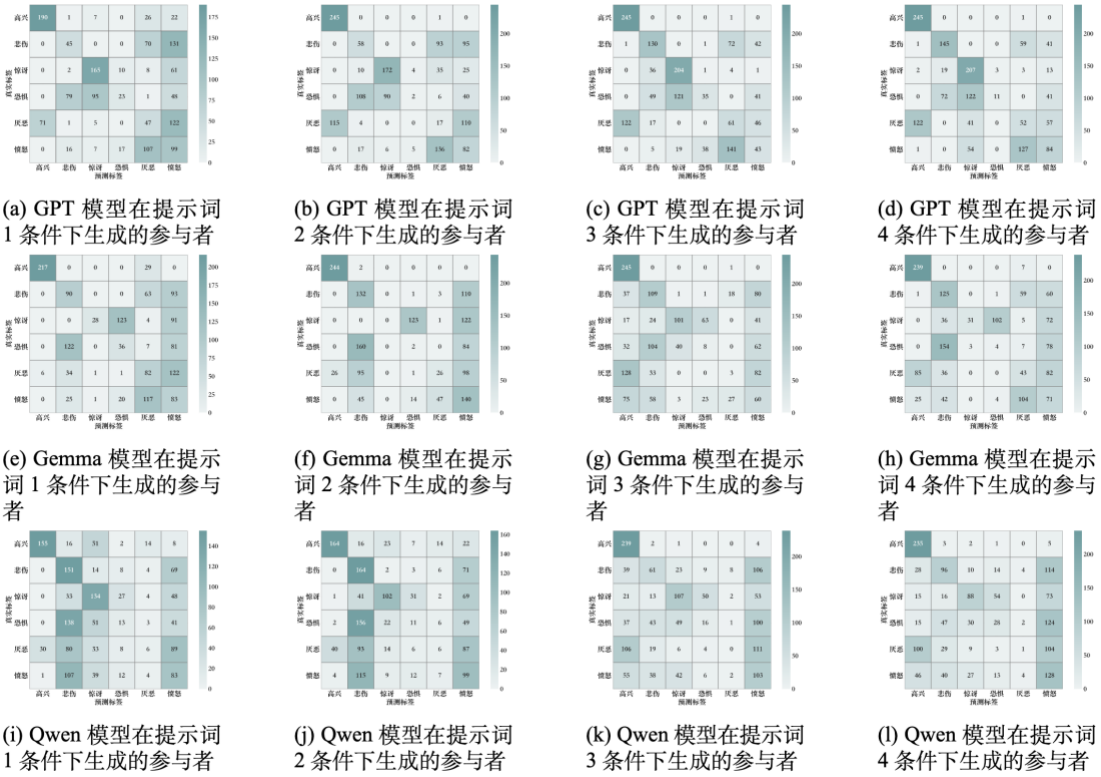
在统计完协调比率以及协调性等级评分后，我们对各模型在四种提示词情况下生成的参与者判断的情绪混淆情况分别进行了统计。接下来依次为 GPT 生成的参与者的情绪混淆情况、Gemma 生成的参与者的情绪混淆情况、Qwen 生成的参与者的情绪混淆情况。

首先，我们对 GPT 模型在四种情况下分别生成的参与者的情绪判断混淆情况进行统计与对比。通过图 10a、图 10b、图 10c、图 10d 得知，GPT 生成的参与者在四种指导语情况下的情绪混淆情况大致相似，即：悲伤容易被误判为厌恶和愤怒，恐惧容易被误判为悲伤和惊讶，厌恶容易被误判为高兴和愤怒，愤怒容易被误判为厌恶。

其次，我们对 Gemma 模型在四种情况下分别生成的参与者的情绪判断混淆情况进行统计与对比。通过图 10e、图 10f、图 10g、图 10h 得知，Gemma 生成的参与者在四种指导语情况下的情绪混淆情况大致相似，即：悲伤容易被误判为愤怒，惊讶容易被误判为恐惧和愤怒，恐惧容易被误判为悲伤和愤怒，厌恶容易被误判为高兴（尤其是在给大模型示例图片之后）、悲伤和愤怒，愤怒容易被误判为厌恶。

最后，我们对 Qwen 模型在四种情况下分别生成的参与者的情绪判断混淆情况进行统计与对比。通过图 10i、图 10j、图 10k、图 10l 得知，Qwen 生成的参与者在提示词有示例图片和无示例图片情况下的情绪混淆情况大致分为两种：1.有示例图片：悲伤容易被误判为愤怒，惊讶被误判为恐惧和愤怒，恐惧被误判为惊讶和愤怒，厌恶被误判为高兴和愤怒，愤怒被误判为高兴；2.无示例图片：恐惧被误判为悲伤，厌恶被误判为高兴、悲伤和愤怒，愤怒被误判为悲伤。

综上所述，不同的 MLLM 具有不同的情绪混淆情况，但仍有一些相似，如：高兴几乎很少有被误判的情况，而悲伤（悲伤容易被误判为愤怒）、惊讶、恐惧（恐惧被误判为愤怒）、厌恶（厌恶容易被误判为高兴和愤怒）和愤怒（愤怒容易被误判为厌恶）均存在误判情况。



(a) GPT 模型在提示词 1 条件下生成的参与者

(b) GPT 模型在提示词 2 条件下生成的参与者

(c) GPT 模型在提示词 3 条件下生成的参与者

(d) GPT 模型在提示词 4 条件下生成的参与者

(e) Gemma 模型在提示词 1 条件下生成的参与者

(f) Gemma 模型在提示词 2 条件下生成的参与者

(g) Gemma 模型在提示词 3 条件下生成的参与者

(h) Gemma 模型在提示词 4 条件下生成的参与者

(i) Qwen 模型在提示词 1 条件下生成的参与者

(j) Qwen 模型在提示词 2 条件下生成的参与者

(k) Qwen 模型在提示词 3 条件下生成的参与者

(l) Qwen 模型在提示词 4 条件下生成的参与者

图 10 MLLMs 在各提示词条件下生成的参与者的情绪混淆矩阵

3.4.4情绪判断依据

由于参与者在判断情绪时会出现情绪判断的混淆情况，所以我们对 12（3*4）种情况下参与者的判断依据进行了统计分析，结果如表 12、

表 **13** 和

表 **14** 所示。

通过方差分析及事后检验，我们得出以下规律：无论是在何种虚拟参与者情况下，高兴都更容易通过下面部进行识别；在 GPT 提示词 3、GPT 提示词 4、Gemma 提示词 3、Qwen 提示词 1、Qwen 提示词 2 这五种情况下的虚拟参与者都更多地通过下面部去识别悲伤情绪；对于惊讶这一情绪来讲，GPT 提示词 1、GPT 提示词 2、GPT 提示词 3、GPT 提示词 4、Qwen 提示词 3、Qwen 提示词 4 这六种情况下的虚拟参与者都是通过上面部进行情绪识别的；对于恐惧这一情绪来讲， GPT 提示词 3 和 GPT 提示词 4 的虚拟参与者是通过上面部进行情绪识别的；对于愤怒这一情绪来讲，Qwen 提示词 3 和 Qwen 提示词 4 的虚拟参与者是通过下面部进行情绪识别的。

通过方差分析及事后检验，我们得出以下规律：无论是在何种参与者情况下，高兴都更容易通过下面部进行识别；在 GPT 提示词 3、GPT 提示词 4、Gemma 提示词 3、Qwen 提示词 1、

Qwen 提示词 2 这五种情况下的参与者都更多地通过下面部去识别悲伤情绪；对于惊讶这一情绪来讲，GPT 提示词 1、GPT 提示词 2、GPT 提示词 3、GPT 提示词 4、Qwen 提示词 3、Qwen 提示词 4 这六种情况下的参与者是通过上面部进行识别的；对于恐惧这一情绪来讲， GPT 提示词 3 和 GPT 提示词 4 的参与者是通过上面部进行识别的；对于愤怒情绪来讲，Qwen 提示词 3 和 Qwen 提示词 4 的参与者是通过下面部进行识别的。

表 12　MLLMs 生成的参与者在各提示词情绪判断任务中的判断依据

参与者		提示词 1		提示词 2		提示词 3		提示词 4	
	情绪	上面部	下面部	上面部	下面部	上面部	下面部	上面部	下面部
GPT	高兴	0%	100%	0.48%	99.52%	2.44%	97.56%	0%	100%
	悲伤	0%	0%	100%	0%	12.82%	87.18%	4.76%	95.24%
	惊讶	100%	0%	100%	0%	95.95%	4.05%	90.91%	9.10%
	恐惧	0%	0%	0%	0%	100%	0%	100%	0%
	厌恶	0%	0%	0%	0%	25%	75%	50%	50%
	愤怒	33.33%	66.67%	0%	100%	100%	0%	81.82%	18.18%
Gemma	高兴	0%	100%	0%	100%	26.70%	73.30%	0%	100%
	悲伤	0%	0%	0%	0%	0%	100%	0%	0%
	惊讶	0%	0%	0%	0%	0%	0%	0%	0%
	恐惧	0%	0%	0%	0%	0%	0%	0%	0%
	厌恶	0%	0%	0%	0%	0%	0%	0%	0%
	愤怒	0%	0%	0%	0%	0%	0%	0%	0%
Qwen	高兴	0%	100%	0%	100%	1.79%	98.21%	0%	100%
	悲伤	1.92%	98.08%	4.76%	95.24%	33.33%	66.67%	47.37%	52.63%
	惊讶	50%	50%	0%	0%	100%	0%	100%	0%
	恐惧	0%	0%	0%	0%	0%	0%	100%	0%
	厌恶	0%	100%	0%	0%	0%	0%	0%	0%
	愤怒	33.33%	66.67%	0%	0%	13.64%	86.36%	0%	100%

表 13 MLLMs 生成的参与者判断依据的混合设计方差分析

变异来源	*df*	*F*	*p*	效应量
参与者类型	2	165.39	<.001	.73
提示词类型	3	44.79	<.001	.27
情绪类型	5	51.88	<.001	.30
参与者类型×提示词类型	6	44.26	<.001	.43
参与者类型×情绪类型	10	26.33	<.001	.31

情绪类型×提示词类型	15	50.52	<.001	.30
参与者类型×提示词类型×情绪类型	30	49.48	<.001	.45

表 14 MLLMs 生成参与者的判断依据（↑ 代表判断依据为上面部，↓ 代表判断依据为下面部)

	GPT				Gemma				Qwen			
情绪	提示词 1	提示词 2	提示词 3	提示词 4	提示词 1	提示词 2	提示词 3	提示词 4	提示词 1	提示词 2	提示词 3	提示词 4
高兴	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
悲伤	-	-	↓	↓	-	↓	-	-	↓	↓	-	-
惊讶	↑	↑	↑	↑	-	-	-	-	-	-	↑	↑
恐惧	-	-	↑	↑	-	-	-	-	-	-	-	-
厌恶	-	-	-	-	-	-	-	-	-	-	-	-
愤怒	-	-	-	↑	-	-	-	-	-	-	↓	↓

3.4.5讨论

在研究二中，我们进一步比较了在不同提示词条件下三种大语言模型生成的参与者在情绪识别任务中的表现。

首先，实验结果揭示了提示词 3（简单文本 + 样例图片）在三个模型上均获得了最优的性能，突出了样例图片对模型理解和推理过程的重要性。提示词 1 和 2 是纯文本的提示词，模型需要将 5 个问题（表 9）中的语言概念映射到视觉特征的理解上，而提示词 3 和 4 中的样例图片为模型提供了具体参照，通过少样本学习（Few shot learning），帮助模型直观理解“协调性”这一抽象概念的具体含义，建立了文本和视觉的语义桥梁，极大程度地降低了任务的模糊性。

这一性能提升也体现在了面部判断依据的准确性，实验结果表明在提示词中引入示例图片引导了模型对面部上下区域的关注策略，例如在 GPT 中，模型更加准确将恐惧归因于上面部。

并且提示词 3 和人类被试的提示词是最为接近的，这也揭示了大模型在理解复杂任务时呈现出与人类相似的学习模式。

与此同时，提示词 2 虽然提供了更加详细、更加具限制性的提示指令，但是识别效果反而更差。这是因为，“请忽略空白分割、灰度分布不均匀、拼图等问题”等指令一方面分散了模型的注意力，另一方面增加了模型的认知负荷，使得模型反而更加关注本该被忽略的线索，从而干扰了本身对于情绪协调性的判断。

三个大模型中，GPT 的性能都相对较优，无论是最差或是最理想的提示词情况，说明了 GPT 的指令理解能力和多模态信息整合能力均更强一些。

除了分析提示词的复杂度，我们进一步基于评分倾向和情绪混淆模式开展讨论。实验结果显示，尽管提示词类型会影响模型的性能表现，但是模型整体的固有的行为模式相对保持稳定。

具体而言，对于评分倾向，虽然样例图片的引入提升了协调比率，但是 GPT 的评分倾向在所有提示词情况下均趋于谨慎；同样的，Gemma 模型低置信度的评分倾向并未发生改变；Qwen 均呈现评分分散且无固定趋势的状态。这说明评分模式主要受到模型内部架构、训练数据和强化学习等因素的影响，临时的任务指令无法左右模型固有的决策风格。

对于情绪混淆，某些固有的情绪混淆模式，例如高兴很少被误判、恐惧误判为惊讶等，在不同指导语的情况下仍然非常稳定，这体现了预训练数据对模型的内在影响。与此同时，提示词中样例图片的引入导致模型可能影响上下面部特征的整合模式，以及可能会过于关注样例中的表情表达模式，例如对于嘴部区域的特征的关注（如图 7 所示），一方面降低了悲伤被误判为厌恶的频率（GPT 和 Gemma），另一方面，增加了厌恶被误判为高兴或者愤怒的评率（Qwen）。虽然样例图片的引入对模型的性能造成了轻微的影响，但是整体情绪混淆趋势保持稳定。

综上所述，实验四验证了文章中先前所提出的假设，即提示词的内容可能系统性地调节大语言模型对情绪拼接面孔的情绪识别表现，进而影响其对于情绪表达的加工和判断。

4 总讨论

本研究探究了当前多模态大语言模型在面部表情识别过程中的识别性能以及影响因素，并回答了先前所提出的假设。研究一 a 将多模态大语言模型生成的虚拟参与者的结果与人类进行对比，探究其与人类之间存在的差异；研究一 b 通过面孔身份的变化来考察研究一中所得出的差异是否具有特定性；研究一 c 通过去除面部分割线这一视觉线索来验证 MLLMs 与人类情绪识别能力差异是否会受到视觉线索的影响；研究二则从提示词这一因素为着手点，来探究会对多模态大语言模型面部表情识别能力产生影响的因素。研究中得出的结果可以进一步帮助多模态大语言模型提升面部表情识别能力。

4.1 多模态大语言模型与人类在面部表情识别中的能力差异

MLLMs 和人类在面部表情识别过程中所展现的差异可能源于二者在信息整合策略存在本质差异。人类能够灵活地调动已有知识经验、当前的情境信息以及所呈现的多种视觉线索进行情绪识别，而当前的 MLLMs 目前主要仍依赖于数据集当中已有的统计模式和符号加工。需要注意的是，在研究一的情绪表达拼接图中，MLLMs 可能会同时收到两类信息：一是上下面部情绪线索的一致或不一致，二是由分割线所带来的视觉线索。为了这两类信息对于实验的影响，我们增加了研究一 c，对分割线进行处理。实验结果表明，有无分割线并不会影响人类和 MLLMs 在上下面部来自于不同的情绪类别的面部表情情绪表达拼接图上的差异。这表明，分割线并不是导致 MLLMs 与人类在情绪识别能力上存在差异的主要原因。而这一实验结果也在一定程度上表明，人类和 MLLMs 在面对上、下面部所传达的视觉线索时采用了不同的信息整合策略。

相较于人类能够将丰富的面部信息线索整合为连贯的整体，MLLMs 则更倾向于优先检测

局部冲突或较为显著的视觉线索，从而导致其在需要进行跨区域信息整合和依赖于经验进行解释的任务中展现出局限性。这一结果也进一步提示我们：情绪识别并不只是低层次的信息匹配，而是一种涉及到社会经验、语义知识和情境加工的复杂的认知加工过程。

在大模型层面，模型的结构设计决定了性能的上限。不同的模型训练集、视觉特征提取方式、多模态融合方式、MLLM 推理方式等因素，最终导致了三个 MLLM 在基于人脸协调/互斥呈现的情绪识别任务中不同的性能表现。同时，由于本身实验任务相较于传统的多模态大模型的视觉任务比较复杂而且具有特异性，不是单纯的视觉特征的提取与匹配，从协调/互斥（上下面部分割）的面部表情图像中提取的视觉线索，可能是矛盾的或者模糊的，这增加了模型进行推理和判断的难度。此外，这不是简单的目标识别的任务，也考察了模型对情感的理解能力。不同的是，人类在情绪识别过程中除了依赖于外显的视觉线索外，还会受到认知和社会经验的调节(Adolphs, 2002; Barrett et al., 2011)。即人类个体在情绪识别过程中依赖多模态信息整合和情境化信息加工，个体接收到视觉刺激后，知觉整合机制会将所观察到的面部特征与情境信息、已有经验、语义知识等相联系，进而在心理层面产生相应的情绪类别(Barrett, 2017a; Calvo & Nummenmaa, 2016)。

相比较之下，大语言模型主要是通过大量训练数据习得规律，从而构建情感识别的能力，并且这种情绪识别是浮于表面的识别(Shanahan, 2024)。但是这种能力在本次实验中并没有得到理想的体现。一方面说明了大模型对情感的理解仍然是一个值得深入研究的方向，另一方面说明通用大模型可能不适用于这种复杂设计的情感识别。大语言模型缺乏灵活的情境构建于整合语义的能力，所以其难以在认知层面形成真正的情绪体验(Bommasani, 2021)。这种大语言模型的情绪识别本质上是符号层面的匹配，而并非是建立在内在的语义整合与情境整合基础上的识别。一旦模型遇到面孔信息较多、信息较为混淆的视觉刺激时，变容易出现情绪混淆识别的情况(Sabour et al., 2024)。因此，本研究中模型表现与人类存在较大差异且不稳定的现象一方面揭示了当前通用的多模态大语言模型在复杂、混淆情绪识别任务中的局限性，另一方面又表现出了人类个体在情绪识别中的认知加工、情境整合与社会经验整合等方面的独特优势(Akben et al., 2025; Gaya-Morey et al., 2024; Lake et al., 2017; Lévêque et al., 2022)。在未来也许通过借鉴人类情绪识别的认知机制，进行专有的模型设计，推动情感智能的进一步发展。

需要进一步指出的是，本文关于人类与 MLLMs 在信息整合方式上的差异讨论，属于基于行为结果对信息整合策略的推断，而非对内部机制（如神经或电生理过程）的直接刻画。与认知心理学中大量基于行为指标推断加工策略的研究传统一致，例如在 Navon 范式或复合面孔效应研究中，研究者通常通过行为表现来推断个体在整体或局部信息加工中的优势模式。本研究同样通过拼接面孔任务中的判断模式，分析不同系统在冲突信息条件下的信息整合方式。因此，本文所揭示的差异更应理解为功能层或算法层的表现差异，而非实现层机制之间的直接对应关

系。

4.2 提示词对多模态大语言模型面部表情识别能力的影响

在心理学研究中，提示词与练习试次不仅仅是帮助参与者熟悉实验操作，并且会影响到参与者的心理表征与准备状态。已有研究结果表明，明确的任务提示词能够帮助个体形成特定的心理集合（mental set），进而能够优化个体的注意分配、反应策略以及判断标准(Logan & Gordon, 2001; Meiran, 2010)。与之相似的是，大语言模型在完成实验任务前，如果能够接收到明确的、结构化的提示词以及示例图片，就会减少模型推理的歧义性，进而使得模型能够产生更好的结果。这种对应表明，无论是人类还是大语言模型，在执行任务前的准确信息引导，都会对参与者的预先设想发挥重要作用，并决定之后的认知加工方向和加工效率。

特别是在多模态的复杂任务中，我们通过少样本的样例学习，弥补文本描述的抽象性，高效对齐模型和人类认知，增强模型对任务的理解，从而提升准确率。

与此同时，应该尽量避免复杂的指令描述，避免模型多度关注限制性的指令。如果上述情况不可避免，同时在指令中明确模型应该关注的部分，引导模型学习正确的归因，可能可以避免模型注意力的偏差。

此外，不容忽视的是，虽然不同的提示词设计可以改变模型的表现，但是总体风格仍受到模型本身架构的限制。未来在与大模型交互的过程中，应该优先以简洁明确的文本描述+示例的结构设计指导语，以获得最优性能。同时，结合讨论二中不同大模型的结构差异，根据不同模型的风格，定制化设计相应示例，例如通过提升样例数目和平衡样本情绪类型等手段，补偿模型本身因训练数据和架构设计带来的限制。

4.3 局限与展望

本研究采用实证研究证明了当前多模态大语言模型在面部表情识别方面与人类存在的显著差异，以及提示词在多模态大语言模型面部表情识别过程中的作用。虽然本研究得到了一系列结果，但仍存在一定的不足，对之后关于多模态大语言模型在面部表情识别方面的研究具有一定的启发性意义。

首先，因为本文想要探讨的是当前多模态大语言模型与人类之间的差异以及性能影响因素，且为了保证较高的生态效度，所以研究中并未采用传统的因素设计。未来的研究可以在保证搞生态效度的前提下，通过系统地移除或干扰某一模态线索，来系统地探讨不同模态线索对 MLLMs 和人类情绪识别能力的影响，以及如何通过这些线索的整合来提升 MLLMs 的面孔情绪识别能力，从而弥补这一局限。其次，研究中的刺激材料均为静态图片，并未采用动态面孔图片。未来的研究应逐步摆脱对于静态面孔刺激的依赖性，使用自发、跨文化且具有更高生态效度的动态面孔刺激，更加全面地探讨 MLLMs 与人类在面部表情识别过程中所存在的能力差异。此外，本研究所选取的多模态大语言模型数量相对有限，不同模型架构可能在情绪识别任务中

表现出不同特征，未来研究可以纳入更多模型进行比较。并且，本研究通过重复推理方式模拟模型个体差异，这种方法虽然能够在一定程度上构建“虚拟参与者”，但仍无法完全等同于真实个体之间的认知差异，未来研究可以探索更加精细的模型行为分析方法。再者，本文主要基于行为指标推断信息整合策略，未涉及神经或电生理层面的证据。未来研究可结合脑电（ERP）、功能性磁共振（fMRI）或眼动等方法，从多层次角度进一步探讨人类与人工系统在面孔情绪加工中的差异机制。最后，在提供给 MLLMs 包含示例图片的提示词时，所包含的图片示例较少。在未来的研究中，可以提供更多的示例图片给 MLLMs，以帮助其能够更好地习得其中的规律，表现出更佳类人的情绪识别能力。

5 结论

综上所述，本研究采用面部表情拼接图片来对比三种多模态大语言模型与人类参与者的情绪识别表现，研究发现 MLLMs 具备初步的面孔情绪识别能力，但仍与人类的情绪识别能力之间存在显著差异。相较于人类能够将丰富的信息整合起来进行情绪识别，MLLMs 主要依赖于从大量训练数据集中学习到的特征进行情绪识别，而不能对多线索进行整合。此外，研究还探究了文本提示词和视觉提示图片对于 MLLMs 的面孔情绪识别能力的影响，发现文本提示的细节增加和视觉提示图片的缺失会使 MLLMs 的识别能力下降，这在一定程度上表明模型会受到外在提示的影响。未来的情绪识别研究可以将多模态大语言模型纳入参与者考虑范畴，在提高对于人类情绪识别能力理解的同时，帮助多模态大语言模型更加智能化。

参考文献

Adolphs, R. (2002). Recognizing emotion from facial expressions: Psychological and neurological mechanisms. *Behavioral and Cognitive Neuroscience Reviews*, *1*(1), 21–62. https://doi.org/10.1177/1534582302001001003

Akben, M., Gude, V., & Ajjan, H. (2025). Silicon minds versus human hearts: The wisdom of crowds beats the wisdom of AI in emotion recognition. *arXiv Preprint arXiv:2508.08830*.

Barrett, L. F. (2017a). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Barrett, L. F. (2017b). The theory of constructed emotion: An active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1), 1–23.

Barrett, L. F., Mesquita, B., & Gendron, M. (2011). Context in emotion perception. *Current Directions in Psychological Science*, *20*(5), 286–290.

Bombari, D., Schmid, P. C., Schmid Mast, M., Birri, S., Mast, F. W., & Lobmaier, J. S. (2013). Emotion recognition: The role of featural and configural face information. *Quarterly Journal of Experimental Psychology*, *66*(12), 2426–2442. https://doi.org/10.1080/17470218.2013.789065

Bommasani, R. (2021). On the opportunities and risks of foundation models. *arXiv Preprint arXiv:2108.07258*.

Brooks, J. A., Chikazoe, J., Sadato, N., & Freeman, J. B. (2019). The neural representation of facial-emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences*, *116*(32), 15861–15870. (world). https://doi.org/10.1073/pnas.1816408116

Bruchmann, M., Mertens, L., Schindler, S., & Straube, T. (2023). Potentiated early neural responses to fearful faces

are not driven by specific face parts. *Scientific Reports*, *13*(1), 4613.

Calder, A. J., Young, A. W., Keane, J., & Dean, M. (2000). Configural information in facial expression perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 527–551. https://doi.org/10.1037/0096-1523.26.2.527

Calvo, M. G., & Nummenmaa, L. (2016). Perceptual and affective mechanisms in facial expression recognition: An integrative review. *Cognition and Emotion*, *30*(6), 1081–1106. (world). https://doi.org/10.1080/02699931.2015.1049124

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., & others. (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45.

Chochlakis, G., Pandiyan, N. M., Lerman, K., & Narayanan, S. (2025). Larger language models don't care how you think: Why chain-of-thought prompting fails in subjective tasks. *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Darwin, C., & Darwin, S. F. (1872). *The expression of the emotions in man and animals* (Vol. 3). John Murray London.

Marr, D. (1982). *Vision: A computational investigation into the human representation and ...* https://books.google.co.jp/books?hl=zh-CN&lr=&id=D8XxCwAAQBAJ&oi=fnd&pg=PR7&dq=Vision:+A+Computational+Investigation+into+the+Human+Representation+and+Processing+of+Visual+Information&ots=KJQFL5Y4hy&sig=8-oc_FargzX7oV5pVscTbiRwxWM&redir_esc=y#v=onepage&q=Vision%3A%20A%20Computational%20Investigation%20into%20the%20Human%20Representation%20and%20Processing%20of%20Visual%20Information&f=false

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3

Young, A. W., Hellawell, D., & Hay, D. C. (2013). Configurational information in face perception. *Perception*, *42*(11), 1166–1178. (Sage UK: London, England). https://doi.org/10.1068/p160747n

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Faustmann, L. L., Eckhardt, L., Hamann, P. S., & Altgassen, M. (2022). The effects of separate facial areas on emotion recognition in different adult age groups: A laboratory and a naturalistic study. *Frontiers in Psychology*, *13*, 859464.

Gaya-Morey, F. X., Ramis-Guarinos, S., Manresa-Yee, C., & Buades-Rubio, J. M. (2024). Unveiling the human-like similarities of automatic facial expression recognition: An empirical exploration through explainable ai. *Multimedia Tools and Applications*, *83*(38), 85725–85753.

Gendron, M., Roberson, D., van der Vyver, J. M., & Barrett, L. F. (2014). Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture. *Emotion*, *14*(2), 251.

Jack, R. E., Garrod, O. G. B., & Schyns, P. G. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*(2), 187–192. https://doi.org/10.1016/j.cub.2013.11.064

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Lévêque, L., Villoteau, F., Sampaio, E. V., Perreira Da Silva, M., & Le Callet, P. (2022). Comparing the robustness of humans and deep neural networks on facial expression recognition. *Electronics*, *11*(23), 4030.

Li, C., & Qi, Y. (2025). Toward accurate psychological simulations: Investigating LLMs' responses to personality and cultural variables. *Computers in Human Behavior*, *170*, 108687. https://doi.org/10.1016/j.chb.2025.108687

Lian, Z., Sun, L., Sun, H., Chen, K., Wen, Z., Gu, H., Liu, B., & Tao, J. (2024). Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, *108*, 102367.

Logan, G. D., & Gordon, R. D. (2001). Executive control of visual attention in dual-task situations. *Psychological Review*, *108*(2), 393.

Martinez, L., Falvello, V. B., Aviezer, H., & Todorov, A. (2016). Contributions of facial expressions and body language to the rapid perception of dynamic emotions. *Cognition and Emotion*, *30*(5), 939–952.

Matt, S., Dzhelyova, M., Maillard, L., Lighezzolo-Alnot, J., Rossion, B., & Caharel, S. (2021). The rapid and automatic categorization of facial expression changes in highly variable natural images. *Cortex*, *144*, 168–184.

Mehra, V., Laban, G., & Gunes, H. (2025). Beyond vision: How large language models interpret facial expressions from valence-arousal values. *arXiv Preprint arXiv:2502.06875*.

Meiran, N. (2010). Task switching: Mechanisms underlying rigid vs. Flexible self control. *Self Control in Society, Mind, and Brain*, 202–220.

Murphy, J., Gray, K. L., & Cook, R. (2017). The composite face illusion. *Psychonomic Bulletin & Review*, *24*(2), 245–261.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. *International Conference on Machine Learning*, 8821–8831.

Rossion, B. (2013). The composite face illusion: A whole window into our understanding of holistic face perception. *Visual Cognition*, *21*(2), 139–253. https://doi.org/10.1080/13506285.2013.772929

Russell, J. A. (1994). Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies. *Psychological Bulletin*, *115*(1), 102.

Sabour, S., Liu, S., Zhang, Z., Liu, J. M., Zhou, J., Sunaryo, A. S., Li, J., Lee, T., Mihalcea, R., & Huang, M. (2024). Emobench: Evaluating the emotional intelligence of large language models. *arXiv Preprint arXiv:2402.12071*.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv Preprint arXiv:2402.07927*.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model. *Cognition and Emotion*, *23*(7), 1307–1351.

Schindler, S., & Bublatzky, F. (2020). Attention and emotion: An integrative review of emotional face processing as a function of attention. *Cortex*, *130*, 362–386.

Shanahan, M. (2024). Talking about large language models. *Communications of the ACM*, *67*(2), 68–79.

Sun, J., Dong, T., & Liu, P. (2023). Holistic processing and visual characteristics of regulated and spontaneous expressions. *Journal of Vision*, *23*(3), 6–6.

Tanaka, J. W., & Farah, M. J. (1993). Parts and wholes in face recognition. *The Quarterly Journal of Experimental Psychology*, *46*(2), 225–245.

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, *17*, 18344909231213958.

Wang, Y., Zhong, W., Li, L., Mi, F., Zeng, X., Huang, W., Shang, L., Jiang, X., & Liu, Q. (2023). Aligning large language models with human: A survey. *arXiv Preprint arXiv:2307.12966*.

Wang, Z., Zhang, Q., Zhang, P., Niu, W., Zhang, K., Sankaranarayana, R., Caldwell, S., & Gedeon, T. (2025). Visual

and textual prompts for enhancing emotion recognition in video. *arXiv Preprint arXiv:2504.17224*.

Weidner, E. M., Schindler, S., Grewe, P., Moratti, S., Bien, C. G., & Kissler, J. (2022). Emotion and attention in face processing: Complementary evidence from surface event-related potentials and intracranial amygdala recordings. *Biological Psychology*, *173*, 108399.

Xu, F., Yuan, Y., Zhang, J., & Wang, J. Z. (2023). High-speed joint learning of action units and facial expressions. In *Modeling Visual Aesthetics, Emotion, and Artistic Style* (pp. 105–126). Springer.

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141.

Young, A. W., Hellawell, D., & Hay, D. C. (2013). Configurational information in face perception. *Perception*, *42*(11), 1166–1178. (Sage UK: London, England). https://doi.org/10.1068/p160747n

Zhang, F., Cheng, Z., Deng, C., Li, H., Lian, Z., Chen, Q., Liu, H., Wang, W., Zhang, Y.-F., Zhang, R., & others. (2025). MME-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv Preprint arXiv:2508.09210*.

Zhang, Q., Wang, Z., Zhang, D., Niu, W., Caldwell, S., Gedeon, T., Liu, Y., & Qin, Z. (2024). Visual prompting in llms for enhancing emotion recognition. *arXiv Preprint arXiv:2410.02244*.

Zhang, Z., Peng, L., Pang, T., Han, J., Zhao, H., & Schuller, B. W. (2024). Refashioning emotion recognition modeling: The advent of generalized large models. *IEEE Transactions on Computational Social Systems*, *11*(5), 6690–6704.

Ziereis, A., & Schacht, A. (2024). Additive effects of emotional expression and stimulus size on the perception of genuine and artificial facial expressions: An ERP study. *Scientific Reports*, *14*(1), 5574. https://doi.org/10.1038/s41598-024-55678-2

Differences in Emotion Recognition Capabilities Between Humans and Multimodal Large Language Models for Spliced Faces

ZHAO Lin [1,2] , LI JingTing [1,2] , LIU Ye [1,2] , MA JunChi[1,3] , WANG SuJing[1,2]

([1] *State Key Laboratory of Cognitive Science and Mental Health, Institute of Psychology, Chinese Academy of Sciences, Beijing 100101, China*)

([2] *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)

([3] *Jiangsu University of Science and Technology, Zhenjiang 212100, China*)

Faces serve as crucial mediums for transmitting social information, including emotions. Humans rely on multi-level cognitive mechanisms, such as holistic and local processing, to efficiently and accurately recognize basic facial emotions. In contrast, while Multimodal Large Language Models (MLLMs) integrate visual encoding components with language reasoning mechanisms, their processing strategies fundamentally differ from human perceptual processing. Comparing their emotion recognition capabilities helps elucidate the differences in emotion perception and reasoning strategies between the two. Furthermore, although existing research indicates that text prompts significantly influence MLLM outputs, their specific effects in the context of facial emotion recognition lack systematic examination.

Based on these premises, this study aims to explore the advantages of holistic and local feature processing in facial emotion recognition, and further investigate whether these processing patterns are consistent between humans and MLLM-generated "virtual participants." Across fourexperiments, the results reveal that when recognizing composite emotional faces, MLLMs exhibit a distinct preference for local features compared to humans, characterized by low composite ratios and a tendency to judge the images as mutually exclusive. Additionally, the level of detail in the prompts and the inclusion of example images significantly alter the models' judgment biases and composite ratios. In conclusion, these findings deepen our understanding of the divergent emotion comprehension pathways between humans and artificial intelligence, offering a new theoretical reference for AI applications in emotion recognition and human-computer interaction.

Keywords: facial action units, mutual exclusivity, emotion recognition, multimodal large language models