Emotion-LLaMA：通过指令微调实现多模态情感识别与推理

摘要

准确的情感感知对于人机交互、教育和心理咨询等各种应用至关重要。然而，传统的单模态方法往往无法捕捉真实世界中情感表达的复杂性，这些表达本质上是多模态的。此外，现有的多模态大语言模型（MLLMs）在音频融合和识别微妙面部微表情方面面临挑战。为了解决这些问题，我们引入了 MERR 数据集，其中包含 28,618 个粗粒度和 4,487 个细粒度的标注样本，涵盖多种情感类别。该数据集使模型能够从多样化的场景中学习并推广到实际应用。

此外，我们提出了 Emotion-LLaMA，这是一种通过情感特定编码器无缝集成音频、视觉和文本输入的模型。通过将特征对齐到共享空间，并结合经过指令微调的改进版 LLaMA 模型，Emotion-LLaMA 显著提升了情感识别和推理能力。大量评估结果表明，Emotion-LLaMA 超越了其他 MLLMs，在 EMER 的 Clue Overlap 指标上取得 7.83 分，在 Label Overlap 指标上取得 6.25 分，在 MER2023 挑战中 F1 得分达到 0.9036，并在 DFEW 数据集的零样本评估中取得最高的 UAR（45.59）和 WAR（59.37）表现。

关键词：情感；多模态大模型

1 引言

情感感知在诸如人机交互 [6, 7] 教育辅助 [13] 和心理咨询 [4, 12] 等应用中起着至关重要的作用。虽然单模态方法（包括面部表情识别 [15]、文本情感分析 [16] 和音频情感识别 [11]）已显示出一定的效果，但真实世界中的情感数据往往是多模态的，整合了文本、音频和图像信息。

尽管许多多模态融合方法已经在特征交互和模态补全方面取得了可喜的进展 [9]，但这些方法在知识层面的交互（这一点对于人类情感推理至关重要）上仍然缺乏探索。最近，多模态大语言模型（MLLMs）在视觉-语言理解 [22]、视觉问答 [27] 和视频理解 [2] 等任务中表现出色。然而，在情感识别 任务上，像 GPT-4 with Vision（GPT-4V）这样的模型仍面临两大挑战：无法处理音频输入以及无法识别面部微表情。

我们认为，缺乏专门的多模态情感指令数据集是限制 MLLMs 有效性的主要原因。这些问题源于以往方法无法有效整合音频输入（音频对于捕捉语调和听觉线索至关重要），以及难以识别细微的面部微表情。这些限制导致了模型在真实场景中的表现不佳。

为了解决这些挑战，我们引入了 MERR 数据集，该数据集支持多模态大模型，并通过指令微调使其能够从多样化的场景中学习并推广到真实世界的应用中。同时，我们提出了 Emotion-LLaMA 模型，该模型通过情感特定编码器整合音频、视觉和文本输入。通过采用指令微调，

Emotion-LLaMA 显著提高了情感识别的准确性和情感推理的深度，为多模态情感分析设立了新的基准。大量实验与评估结果表明，Emotion-LLaMA 在 EMER、MER2023 和 DFEW 数据集上取得了顶尖成绩，展现了其卓越性能。

2 相关工作

2.1 多模态大语言模型（MLLMs）

多模态大语言模型（MLLMs）[1] 因其强大的推理能力而受到广泛关注。研究主要集中在利用预训练模型（如 CLIP [25]、Q-Former [17] 和 ImageBind [10]）来实现通用领域的应用 。然而，即使是像 GPT-4V [20] 这样的先进模型，也由于缺乏在多模态情感数据集和情感相关知识上的专业训练，难以理解音频中的情感线索和识别面部微表情。最近，研究人员开始在多模态情感数据集上训练 MLLMs，以识别对话中触发情感的语句，但这些研究往往缺乏详细的解释。相比之下，我们提出的 Emotion-LLaMA 采用情感专属编码器来提取多模态特征，从而增强了情感识别和推理能力。

2.2 指令微调

语言指令已被广泛应用于各种 NLP 任务 [3]。研究如 InstructionGPT [24]、FLAN [8] 和 OPT-IML [14] 探索了指令微调方法，显著增强了大语言模型的零样本和少样本能力。视觉领域也采用了语言指令用于各种视觉-语言任务。例如，LLaVA [23] 使用纯语言模型将图文对转换为指令遵循数据，EmoVIT [26] 通过配对注释生成了视觉情感指令数据。然而，这些方法通常缺乏音频信息，而音频信息对理解人类情感至关重要。由于高昂的标注成本，AffectGPT [21] 仅手动标注了 100 个包含情感线索的样本。为了解决情感相关指令遵循数据稀缺的问题，我们的方法利用先验知识生成了多模态描述数据。

3 本文方法

3.1 MERR 数据集的构建

此部分对本文将要复现的工作进行概述，图的插入如图 1 所示：MERR 数据集是通过在视频数据中进行情感表达注释的综合过程构建的，如算法 1 和图 1 所示。MERR 数据集通过一个全面的情感表达标注流程构建。首先，利用 OpenFace 工具从每一帧视频中提取面部特征，该工具检测并评分动作单元（Action Units, AUs），以确定累计强度最大的帧：

$$I_{\text{peak}} = \arg\max_{k} \left(\sum_{i} S_{a_{ui}}^{k} \right), \tag{1}$$

其中 $S_{a_{ui}}$ 表示每个 AU 的强度。这些 AU 被映射为面部表情描述 C_{ved}，以准确反映面部动作。接下来，MiniGPT-v2 分析峰值帧以提取上下文信息 C_{vod}，例如活动和环境，从而帮助识别背景中的潜在情感元素。Qwen-Audio 处理音频片段，提取语音和语调中的细微差异，生成与情感相关的描述 C_{atd}。视觉和音频信息被整合为原始多模态描述，结合感官输入以增强上下文补充，同时还包含文本字幕 C_{ls}，提供与音频和视觉数据互补的文本上下文。LLAMA

将单模态描述（$C_{\text{ved}}, C_{\text{vod}}, C_{\text{atd}}, C_{\text{ls}}$）聚合为详细的多模态描述 C_{md}。最后，综合描述 C_{md} 被用于标注峰值帧，确保视频被详细的情感描述符标注。

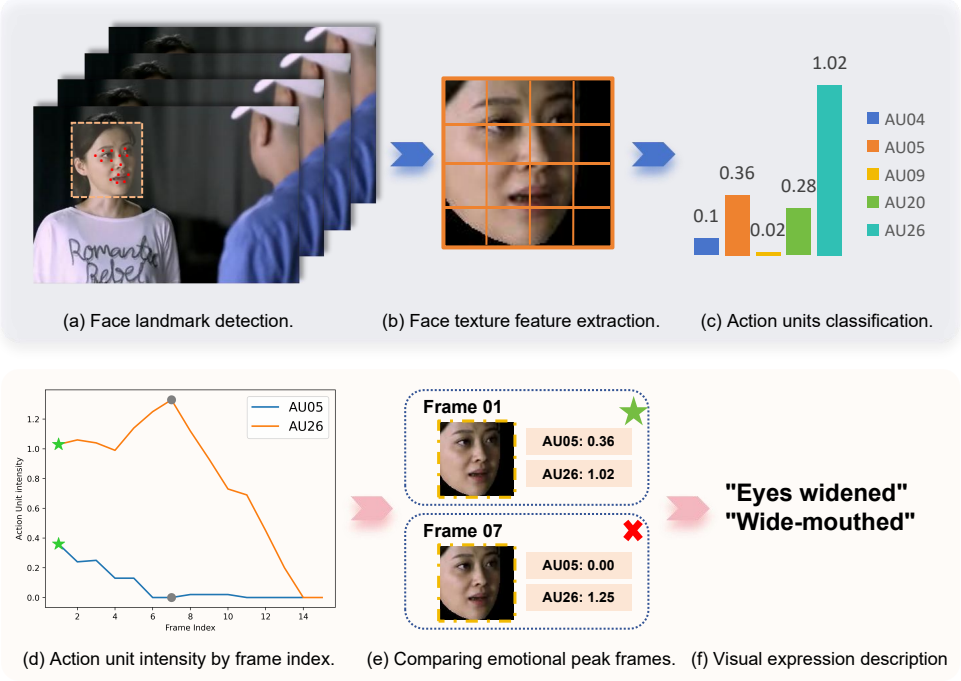

图 1. 峰值帧检测示意图

MERR 数据集扩展了情感类别和注释的范围，超出了现有数据集中的范围. 每个样本都有一个情感标签，并根据其情感表达进行描述. 该数据集最初使用粗粒度标签对来自大量未注释数据池的 28618 个样本进行了自动注释，后来经过改进，包括 4487 个具有细粒度注释的样本. 与其他数据集相比，MERR 包含了更广泛的情感类别，包括“怀疑”和“蔑视”等容易混淆的类别。

3.2 多模态 Emotion-LLaMA 模型

多模态情感大语言模型 (Emotion-LLaMA) 架构如图 2 所示，包含一个音频编码器 E_{aud}、一个视觉编码器 E_{vis} 和一个多模态大语言模型 ϕ。给定输入元组 $P = \langle$音频, 视频, 提示词$\rangle$，Emotion-LLaMA 的公式定义为：

$$\hat{O} = \Psi(\phi, E, \Omega, P) = \phi(E_{\text{aud}}(\text{音频}), E_{\text{vis}}(\Omega(\text{视频})), E_{\text{tex}}(\text{提示词})) \tag{2}$$

其中，ϕ、Ω 和 E 分别表示 LLaMA 语言模型、视觉预处理器和多模态编码器。$\hat{O}$ 表示格式化的输出文本结果。多模态编码器 E 包括音频、视觉和文本提示词编码器。输入 视频 被预处理以构造帧序列 V 和 峰值帧。

3.3 多模态提示模板

为了解决情感理解的复杂需求，我们设计了一个结构化的多模态提示模板，该模板包含描述性标题和情感标记，指导 LLM 解读情感状态与对应视觉或听觉内容之间的潜在关联。模板表示为：

[INST]⟨音频特征⟩⟨视频特征⟩[任务标识]提示词[/INST]

3.4 多视角多模态编码器

为了捕捉音频和视觉模态中的情感线索，我们利用 HuBERT 模型作为音频编码器 E_{aud}，并使用一个多视角视觉编码器 E_{vis}。

HuBERT 从输入音频信号 A 中提取全面的听觉表示 u_{aud}，在情感识别任务中表现出色。

我们使用视觉预处理器统一视觉模态，包括从输入视频中提取的面部序列和峰值帧。三个视觉编码器 $E_{\text{vis}} = \{E_{\text{vis}}^{\text{glo}}, E_{\text{vis}}^{\text{loc}}, E_{\text{vis}}^{\text{temp}}\}$ 被用来全面提取互补的多视角视觉情感特征：

- **局部编码器：** 一个基于 ViT 的模型，使用 MAE 方法 预训练，提取静态面部表情特征。面部序列被输入到局部编码器中，输出的逐帧特征通过平均池化融合，得到局部视觉特征：

$$u_{\text{vis}}^{\text{loc}} = \text{AVG}(E_{\text{vis}}^{\text{loc}}(V))$$

- **时间编码器：** 一个 VideoMAE 模型，生成时间特征：

$$u_{\text{vis}}^{\text{temp}} = E_{\text{vis}}^{\text{temp}}(V)$$

表示面部序列，学习指示情感状态的面部动态，并提供人类情感的时间动态视图。

- **全局编码器：** 一个基于 ViT 的模型 EVA，使用官方预训练权重初始化，生成视觉特征：

$$u_{\text{vis}}^{\text{glo}} = E_{\text{vis}}^{\text{glo}}(\text{峰值帧})$$

捕捉面部表情及背景环境。

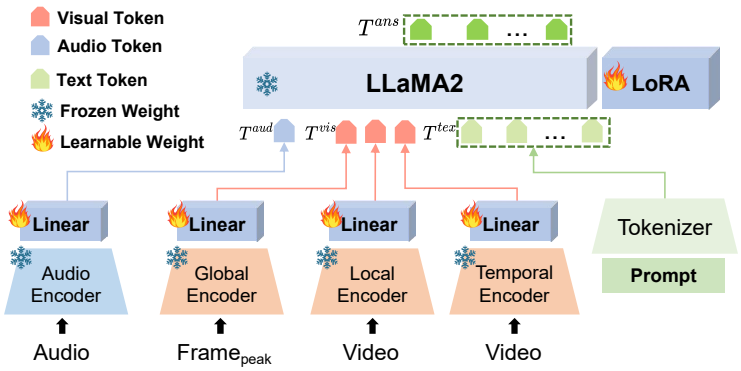

图 2. Emotion LLaMA 的架构，它集成了音频、视觉和文本输入，用于多模态情感识别和推理。

3.5 多模态融合与分词

为了促进对文本输入的高效处理，针对多模态情感推理，我们提出了一种改进的生成方法。该方法通过迭代选择最可能的标记，生成具有上下文意义和情感细腻的响应。

为了将音频和视觉特征与文本标记融合，我们引入了一种线性投影机制，该机制将这些特征转化为统一的维度空间。这涉及可训练的线性映射 σ，包括用于音频标记的 σ_{aud} 和用于视觉标记的 $\sigma_{\text{vis}}^{\text{glo}}, \sigma_{\text{vis}}^{\text{loc}}, \sigma_{\text{vis}}^{\text{temp}}$。具体来说，我们通过 σ 将多模态特征 u 转换为语言嵌入标记 T：

$$T = \sigma \cdot u, \quad u = \{u_{\text{aud}}, u_{\text{vis}}^{\text{glo}}, u_{\text{vis}}^{\text{loc}}, u_{\text{vis}}^{\text{temp}}\} \tag{3}$$

生成的多模态标记 T 包括单个音频标记 $\langle T_{\text{aud}} \rangle$，三个视觉标记 $\langle T_{\text{vis}}^{\text{glo}} \rangle, \langle T_{\text{vis}}^{\text{loc}} \rangle, \langle T_{\text{vis}}^{\text{temp}} \rangle$，以及一系列文本标记 $\langle T_{\text{tex}}^{0} \rangle, \ldots, \langle T_{\text{tex}}^{N} \rangle$。这些标记通过 Emotion-LLaMA 的内置交叉注意力机制进行融合，使模型能够捕捉并推理多模态输入中的情感内容。

通过这种线性投影和多模态标记表示，Emotion-LLaMA 能够处理并整合来自各种模态的信息，既发挥了底层 LLaMA 模型的优势，又融入了音频和视觉来源的关键情感线索。

3.6 Emotion-LLaMA 模型的训练

Emotion-LLaMA 的训练采用从粗到细的方式，分为两个阶段：预训练阶段和多模态指令微调阶段。

阶段 1：预训练

最初，模型在粗粒度数据上进行训练，包括视觉和音频描述。不同的任务标识符帮助模型从多个视角掌握情感。此阶段涉及简单的描述或分类任务，促进多模态特征标记（$\langle T_{\text{aud}} \rangle$, $\langle T_{\text{vis}}^{\text{glo}} \rangle, \langle T_{\text{vis}}^{\text{loc}} \rangle, \langle T_{\text{vis}}^{\text{temp}} \rangle$）与词嵌入空间的快速对齐。

阶段 2：多模态指令微调

在预训练的 Emotion-LLaMA 模型基础上，我们利用精细化指令数据集进一步微调模型，以增强其情感识别和推理能力。此阶段使用多模态指令微调数据集，结合来自 MERR 数据集的综合推理描述。微调过程扩展到多个来源，包括 MER2023 和 DFEW ，这些数据集具有精确注释的情感类别。此阶段确保模型不仅能够准确识别情感，还能理解每种情感背后的上下文和推理过程。

4 与已有开源代码对比

在本研究中，我们利用 Emotion-LLaMA [5] 的开源部分作为基础，但发现其缺少局部编码器、时间编码器和全局编码器的实现。因此，我们参考 Mertools [19] 的相关方法，补充了这些模块的代码。此外，我们对模型的整体架构进行了优化，使用 Video-llama2 框架进行了重新训练，同时引入了 AffectGPT [18] 中的数据，进一步增强了模型在情感分析任务中的性能。

5 实验环境搭建

为了复现实验结果，我们搭建了三个独立的运行环境：`Emotion-LLaMA` 、`Mertools` 和 `Video-llama2`。实验运行硬件配置为 8 × NVIDIA A100 GPU（每张 GPU 配备 40 GB HBM2 显存），并采用 NVIDIA CUDA Toolkit 和 CUDNN 加速深度学习模型的训练和推理。系统运行在 Ubuntu 20.04 操作系统下，依托于具备高性能计算能力的服务器，搭载 AMD EPYC 7742 处理器，提供多核支持以优化数据加载和模型并行性能。此外，存储设备使用 NVMe SSD，确保大规模数据集的快速读取和写入，以满足深度学习任务对计算和存储的高要求。

5.1 Mertools 环境

Mertools 环境的配置文件可以通过以下链接获取：

- https://github.com/zeroQiaoba/MERTools/blob/master/environment.yml

5.2 Video-llama2 环境

Video-llama2 环境的依赖文件可通过以下链接查阅：

- https://github.com/DAMO-NLP-SG/VideoLLaMA2/blob/main/requirements.txt

5.3 Emotion-LLaMA 环境

Emotion-LLaMA 环境的详细配置文件可以访问以下链接：

- https://github.com/ZebangCheng/Emotion-LLaMA/blob/main/environment.yml

5.4 LLaMA2-Chat-7B 模型

LLaMA2-Chat-7B 模型的权重可以通过以下链接获取：

- https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

5.5 Video-llama2-AV-7B 模型

Video-llama2-AV-7B 模型的权重可通过以下链接获取：

- https://huggingface.co/DAMO-NLP-SG/VideoLLaMA2.1-7B-AV

6 数据集获取

6.1 MER2023 数据集

MER2023 数据集的详细说明以及获取方式可以访问以下链接：

- http://www.merchallenge.cn/workshop

6.2 EMER 数据集

EMER 数据集的详细信息和获取方法可以通过以下链接查阅：

- https://github.com/zeroQiaoba/AffectGPT

7 数据集构建细节

7.1 MERR 数据集构建

MERR 数据集来源于 MER2023-SEMI，如图 3 所示，其包含超过 70,000 个未标记的视频剪辑。我们利用几个强大的多模态模型从不同的模态中提取情感线索，然后使用最新的 LLaMA-3 模型总结所有情感线索进行推理，从而得出最终的多模态描述。

图 3. 数据集示例图

7.2 数据过滤

我们使用 OpenFace 从视频片段中提取人脸，然后对齐以识别各种面部肌肉运动，从而检测动作单元。这些肌肉运动的某些组合与特定情绪相关。例如，惊讶的情绪是通过 Action Unit 05（上眼睑提升器）和 26（下巴下垂）的组合来识别的。行动单元的每个特定组合都被分配了一个伪标签，表示样本被选中并表现出强烈的情绪表达特征。总共选择了 28,618 个样本并分配了伪标签。

7.3 视觉表情描述

由于视频中的自然动作（例如眨眼和说话），因此从不同的帧中提取了动作单元（AU）的不同组合。因此，确定最准确地代表当前情绪的 AU 至关重要。我们的方法涉及分析动作单元的振幅值，以确定情绪表达的峰值，称为 “情绪峰值帧”。具体步骤包括：（1）确定所有帧中出现频率最高的动作单元;（2）对它们的值求和，最高的总数表示情绪峰值帧。然后将该帧的动作单元映射到其相应的视觉表情描述。在使用 OpenFace 提取特征后，生成四个文件夹，分别存储不同类型的数据：**openface_all** 包含从输入视频或人脸目录提取的所有特征，包括动作单元（AUs）、面部关键点、姿态和 HOG 特征，便于全面分析；**openface_hog** 专用于存储 HOG（梯度方向直方图）特征，捕捉图像局部纹理信息，常用于物体检测和人脸识别，如图 5 所示；**openface_pose** 保存头部姿态特征，包括旋转角度和三维位置，用于建模人物头部姿势和空间信息，如图 6；**openface_face** 存储对齐后的人脸图像，为表情识别或其他特征分析提供标准化输入，如图 4。这种结构高效管理特征数据，支持多种情感和行为分析任务。

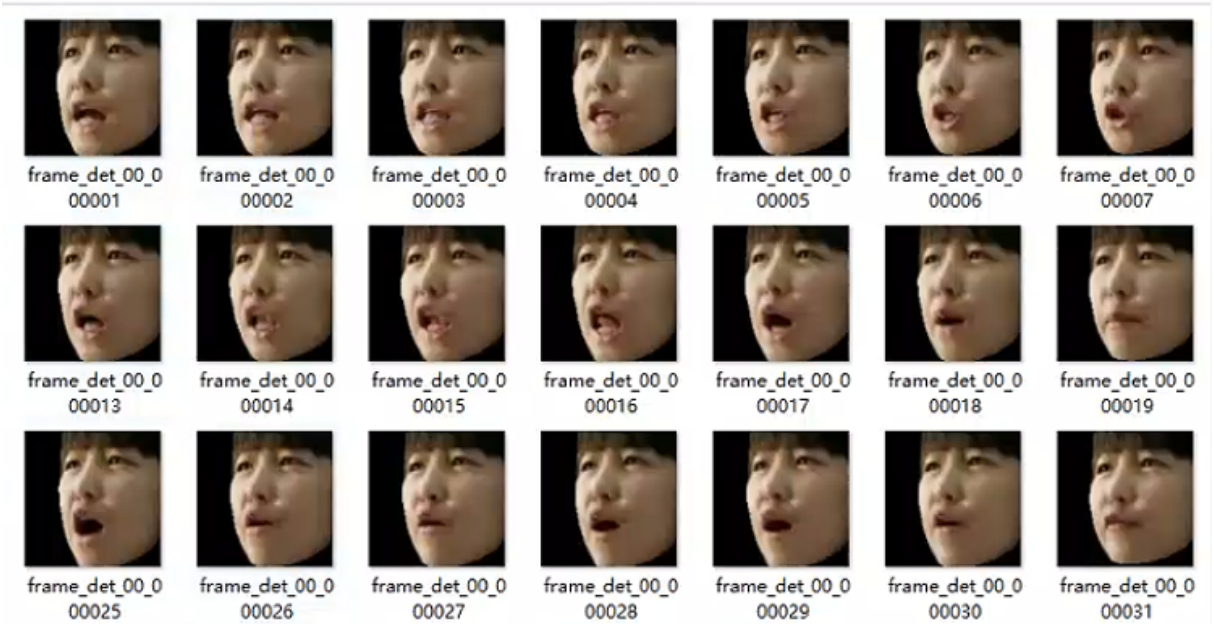

图 4. openface_face

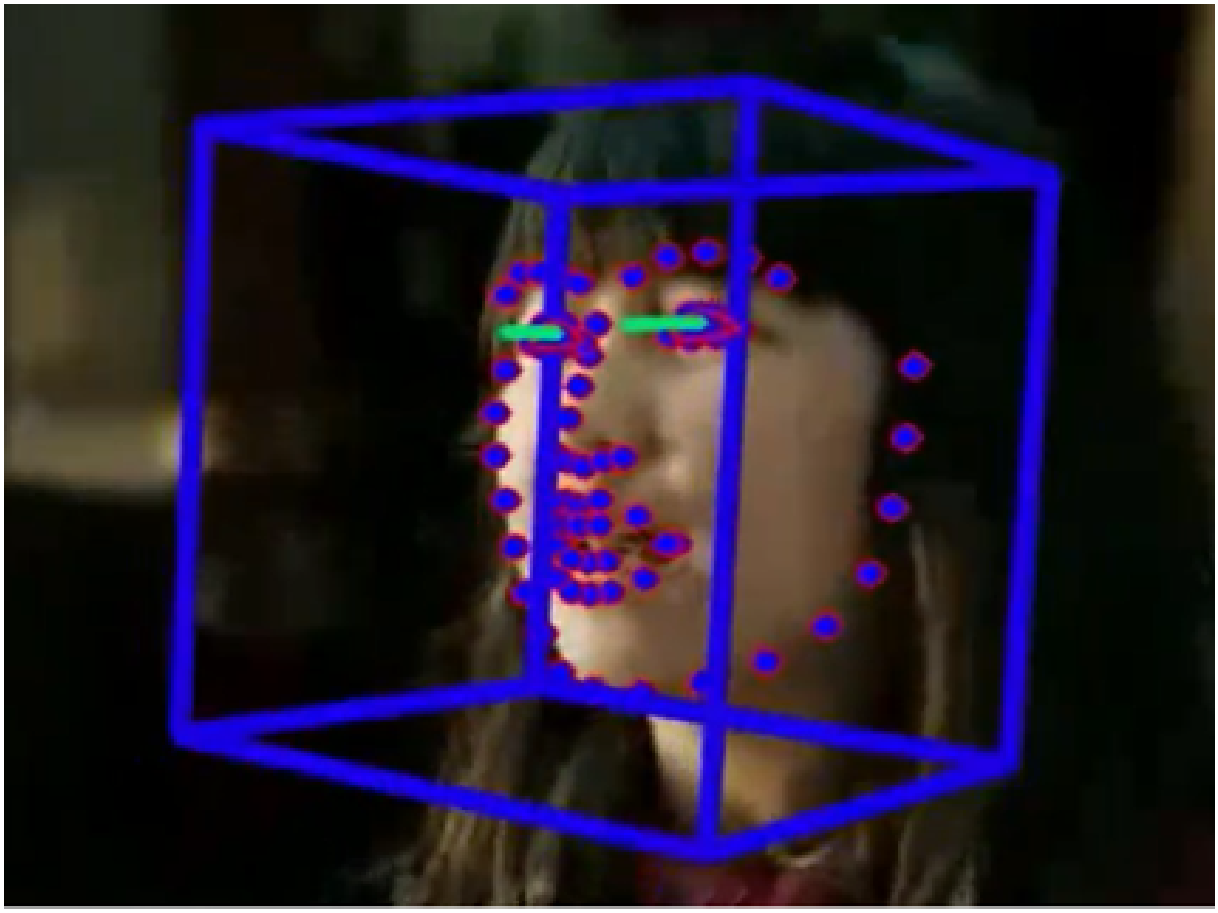

图 5. openface_hog

frame, face_id, timestamp, confidence, success, gaze_0_x, gaze_0_y, gaze_0_z, gaze_1_x, gaze_1_y, gaze_1_z, gaze_angle_x, gaze_angle_y, eye_lmk_x_0, eye_lmk_x_1, eye_lmk_x_2, eye_lmk_x_3, eye_lmk_x_4, eye_lmk_x_5, eye

1, 0, 0.000, 0.98, 1, -0.592586, 0.198135, -0.780759, -0.597904, 0.282822, -0.750015, -0.661, 0.304, 747.6, 749.7, 753.7, 757.4, 758.6, 757.1, 752.5, 748.8, 746.2, 746.8, 748.6, 752.6, 758.1, 762.6, 765.9, 761.7, 756.

2, 0, 0.042, 0.98, 1, -0.537347, 0.216370, -0.815133, -0.602706, 0.277284, -0.748237, -0.630, 0.306, 748.0, 750.6, 755.8, 760.4, 761.9, 759.8, 754.1, 749.5, 745.2, 746.3, 748.2, 753.2, 758.8, 763.3, 766.6, 762.3, 757.

3, 0, 0.083, 0.98, 1, -0.547071, 0.222494, -0.806976, -0.603166, 0.272693, -0.749553, -0.636, 0.308, 747.7, 750.3, 755.3, 759.9, 761.3, 759.2, 753.6, 749.1, 745.1, 746.2, 748.8, 752.8, 758.5, 763.0, 766.3, 762.0, 756.

4, 0, 0.125, 0.98, 1, -0.541515, 0.219738, -0.811466, -0.596884, 0.279851, -0.751940, -0.629, 0.309, 747.9, 750.5, 755.6, 760.2, 761.6, 759.5, 753.9, 749.3, 745.2, 746.2, 748.6, 753.0, 758.7, 763.2, 766.5, 762.2, 756.

5, 0, 0.167, 0.98, 1, -0.542395, 0.243500, -0.804062, -0.578170, 0.258134, -0.774007, -0.617, 0.308, 748.6, 751.2, 756.2, 760.6, 761.9, 759.8, 754.3, 749.9, 745.6, 746.9, 749.4, 753.6, 759.0, 763.4, 766.8, 762.5, 757

6, 0, 0.208, 0.98, 1, -0.573207, 0.260955, -0.776748, -0.731867, 0.277081, -0.622573, -0.751, 0.367, 753.2, 755.2, 759.2, 762.8, 763.9, 762.3, 757.9, 754.3, 750.7, 751.5, 753.5, 757.2, 761.9, 765.8, 768.8, 765.1, 760.

7, 0, 0.250, 0.98, 1, -0.533651, 0.268519, -0.801944, -0.734962, 0.280450, -0.617396, -0.729, 0.369, 753.2, 755.1, 759.3, 763.3, 764.8, 763.3, 758.7, 754.7, 749.4, 750.7, 753.1, 757.0, 761.8, 765.8, 768.8, 765.2, 760.

8, 0, 0.292, 0.98, 1, -0.522026, 0.239791, -0.820788, -0.664071, 0.184514, -0.724564, -0.655, 0.362, 752.8, 754.5, 758.8, 763.2, 765.0, 763.0, 759.1, 754.7, 749.4, 750.4, 753.0, 755.3, 760.3, 765.0, 768.8, 765.0, 759.

图 6. openface_pose

7.4 视觉目标描述

我们将完整的情感峰值帧输入到 MiniGPT-v2 模型中，使其能够描述视频的场景、角色手势和其他方面。

7.5 音频音调描述

我们使用音频作为 Qwen-Audio 模型的输入，然后描述说话者的语气和语调，从而产生对理解情绪同样重要的音频线索。

7.6 粗粒度合成

通过将视觉和音频描述与模板化序列中的词汇字幕集成，我们生成了粗粒度的情感描述。总共制作了 28,618 个这样的描述，如图 7 所示。


```
{
    "sample_00000003": {
        "AU_list": [
            "AU12",
            "AU14"
        ],
        "visual_prior_list": [
            "Toothy smile",
            "Indentation when smiling"
        ],
        "audio_prior_list": "speaks in a normal tone.",
        "peak_index": "12",
        "peak_AU_list": [
            "AU04",
            "AU10",
            "AU12",
            "AU14",
            "AU17",
            "AU23"
        ],
        "pseu_emotion": "happy",
        "caption": "Toothy smile, Indentation when smiling, speaks in a normal tone."
    },
```


图 7. 粗粒度数据集示例图

7.7 细粒度生成

仅仅将组件连接起来并不能真正解释情绪背后的触发因素。因此，我们将所有情绪线索输入到 LLaMA-3 模型中，筛选并正确识别相关线索，结合不同的线索进行推理，从而得出全面的情绪描述，如图 8 所示。

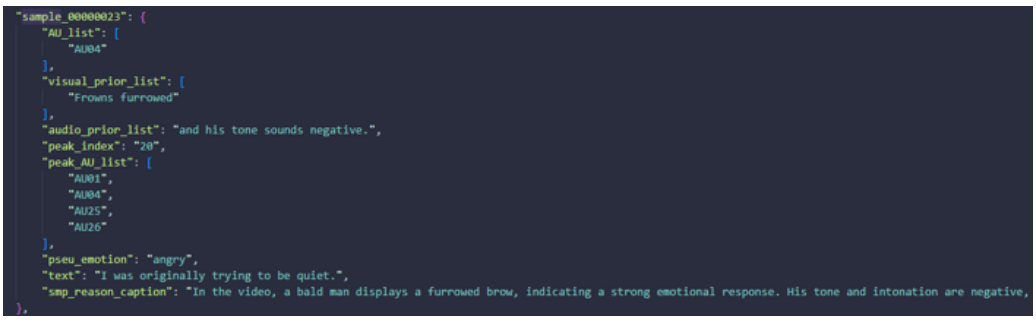

图 8. 细粒度数据示例图

8 合成数据集的 Prompt 设计

为了生成用于情感分析的高质量合成数据集，我们设计了以下 Prompt，用于指导不同模型生成情感描述。

8.1 MiniGPT-v2 的 Prompt

以下是 MiniGPT-v2 模型的 Prompt 设计，用于分析视觉场景：

"What do you see in the image? Please describe it using words only."

"Please describe the content of the image."

这些 Prompt 的设计目的是帮助模型准确描述图像中的场景、角色动作和其他关键视觉信息。

8.2 LLaMA-3 的 Prompt

以下是用于 LLaMA-3 模型的 Prompt，重点在于情感特征的推理：

System: You are an emotion analysis expert. Please infer emotion label based on the given emotional features.

Question: The woman in the video is talking to a man, possibly discussing something important or sharing her thoughts and feelings. The personś expression and action include eyes widened, wide-mouthed, and speaking with a happy voice, saying: Öh my.Please sort out the correct emotional clues and infer why the person in the video feels surprise.

该设计目的是引导模型结合视觉和语境信息，准确识别和推理情绪标签。

8.3 Qwen-Audio 的 Prompt

针对音频数据的情感分析，我们设计了以下 Prompt：

'Does this person in audio speak with a [angry, cheerful, crying, laughter, heartbroken, whispery, trembling, skeptical, astonished, repulsed, sorrowful] tone?'

该 Prompt 强调了语气和语调的重要性，帮助模型生成与音频情感相关的线索。

9 模型的复现细节

9.1 特征提取

在本研究中，特征提取过程基于分组化的层级文件结构进行，以确保数据的分类管理和处理的规范性。特征文件被组织存储于不同的子文件夹中，例如 `features_of_MER2023-SEMI` 和 `HL-UTT`，分别对应 MER2023 半监督学习相关特征和高层语音单元特征（UTT）。特定文件夹如 `mae_340_UTT` 和 `maeV_399_UTT` 则存储了从原始数据中提取的单元特征，编号可能与实验条件或模型配置相关。此外，帧级特征处理文件夹 `first_frames_MER2023-SEMI` 专注于 MER2023 数据集中初始帧的特征提取，适用于时间敏感任务如情感分析。同时，文件 `relative_test3_NCEV.txt` 用作基准测试，用于评估特征提取的性能和实验的可重复性，如图 9 所示。

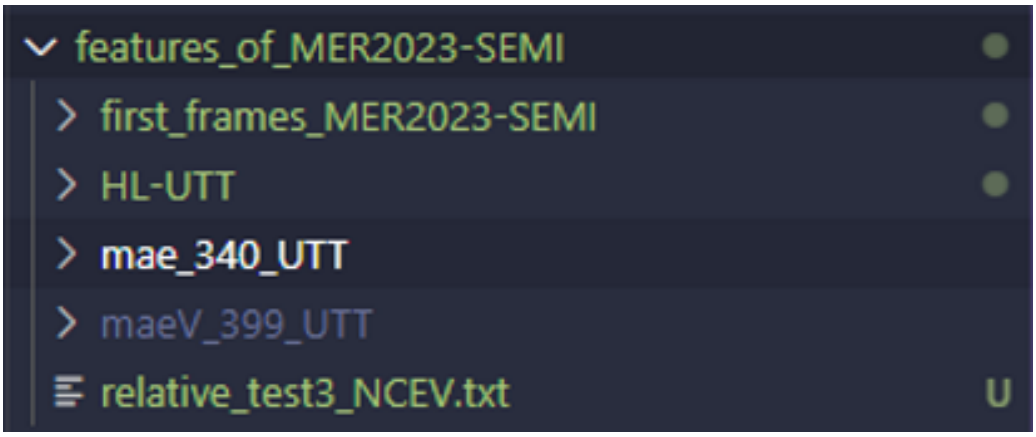

图 9. 提取特征展示图

9.2 大语言模型

从 huggingface 上下载 LLaMA-2-7B-Chat，如图 10 所示

图 10. 下载大模型展示图

9.3 在 MER2023 数据集上复现效果复现效果

复现结果与论文结果相同，Accuracy 为 0.9017，Precision 为 0.8999，Recall 为 0.9017，F1 Score (F1 值): 0.8999，如果 11 所示。

10 基于 Video-llama2 构架的优化细节

尽管 Emotion-LLaMA 在 MER2023 情感分类数据集上取得了较高的分类性能，但其分类过程依赖于预先提取的特征，难以实现端到端的情感分类。而基于 Video-llama2 的架构则具备

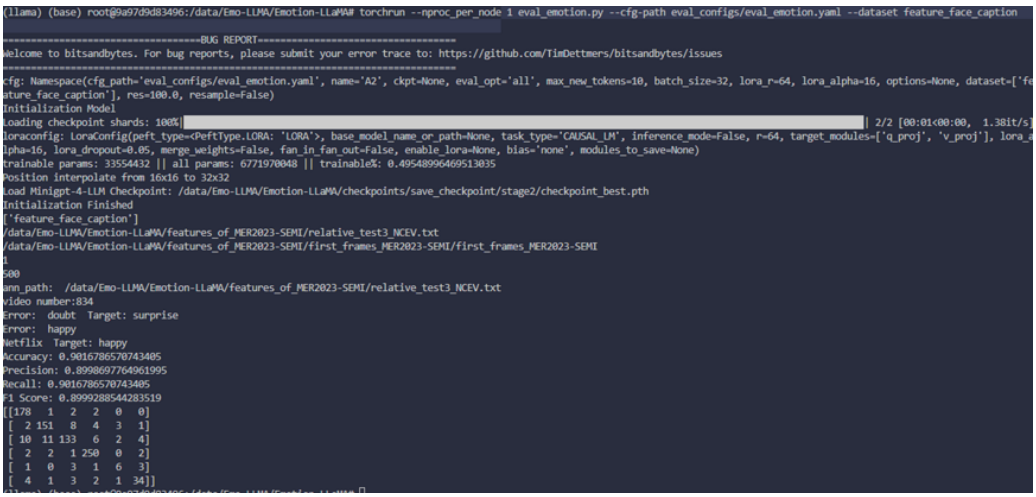

图 11. 提取特征展示图

端到端情感分类的能力，不再依赖特征的独立提取，Video-llama2 的架构如图 12 所示。此外，通过进一步扩展数据集的规模与种类，系统性能和适用性有望得到显著提升，VideoLLama2 情感分析界面示例如图 13 所示。

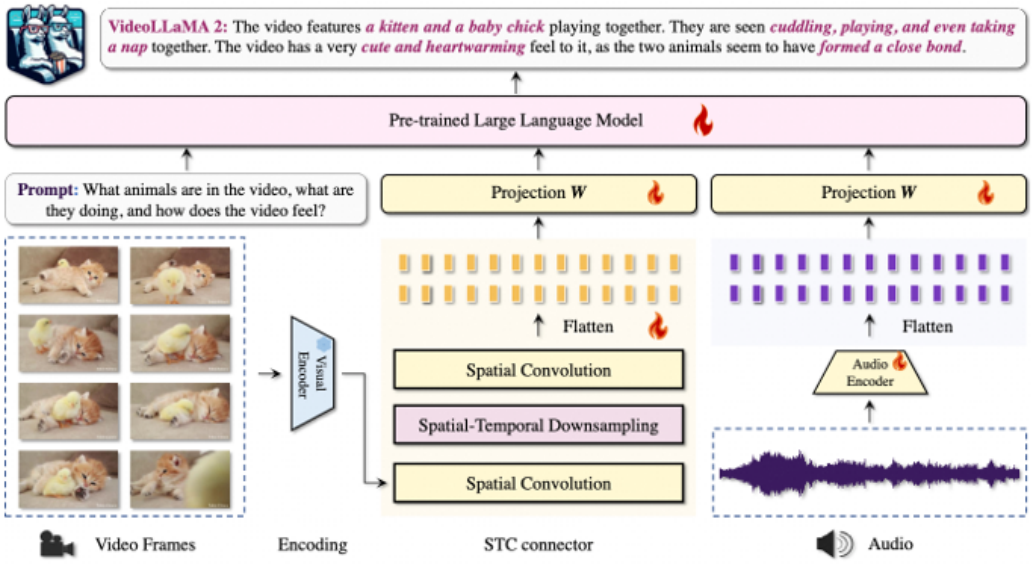

图 12. 提取特征展示图

在 Video-llama2 的训练过程中，采用了两阶段的训练策略：

- **阶段 1**：使用 Emotion-LLaMA 生成的伪标签数据和 AffectGPT 提供的多模态情感分类数据进行初步训练，以增强模型对多模态特征的理解和情感分类的泛化能力。
- **阶段 2**：基于 MER2024 数据集的标注训练数据进一步微调模型，从而提升模型在目标任务上的性能和准确性。

实验结果表明，采用两阶段训练策略能够显著提升模型的情感分类性能。在第一阶段，基于伪标签和多模态情感分类数据的初步训练使模型具备了较好的基础分类能力，测试准确率达到 79.62%，但精确率和召回率存在一定的不平衡，F1 分数仅为 0.77。通过第二阶段引入

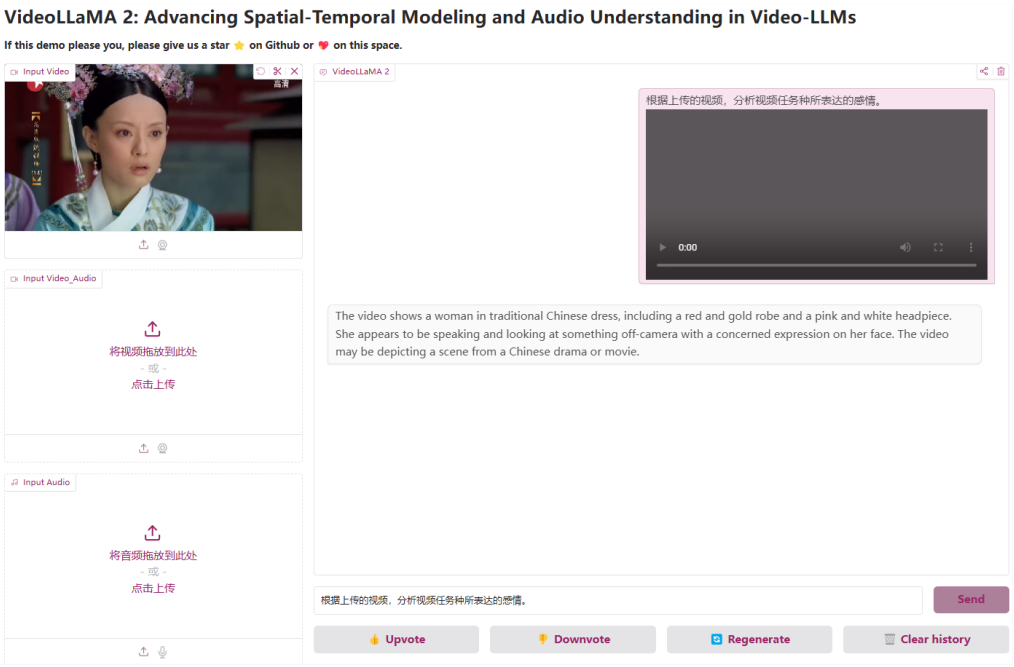

图 13. VideoLLama2 情感分析界面示例图

MER2024 标注数据对模型进行微调后，测试准确率提升至 85.25%，同时精确率、召回率和 F1 分数均达到 0.85，模型的分类性能得到了全面增强。实验结果验证了标注数据在提升模型性能中的关键作用，并表明两阶段训练策略在多模态情感分类任务中具有良好的应用潜力。

11 总结与展望

在多模态情感分类任务中，尽管 Emotion-LLaMA 模型取得了较高的准确率，但其端到端的实际应用方案仍然存在一定局限。未来的研究将聚焦于开发更加适用于实际场景的大规模端到端情感分析模型，以提升模型的实用性和落地能力。

目前的情感大模型研究主要集中于单一情感识别任务。然而，随着情感理解需求的多样化，未来的研究计划将扩展任务范围，涵盖幽默、讽刺等复杂情感表达的识别，并进一步探讨笑声背后的心理和社会机制。这不仅将丰富情感大模型的任务类型，还将拓展其应用场景，从而为更全面的情感分析提供坚实的基础。

在多模态情感大模型的评估领域，现阶段尚缺乏统一的标准，评估方法呈现出多样化和分散化的特点。为了推动该领域的进一步发展，未来的研究将致力于构建一个统一的多模态情感评估基准（benchmark），以标准化评估流程，确保评估的客观性和一致性，从而为模型性能的全面对比和优化提供有力支撑。

虽然 Video-LLaMA2 在多模态情感分类任务中展现了出色的表现，但仍面临诸多挑战。例如，该模型对高质量标注数据的依赖性较强，多模态特征融合能力有待进一步提升，且计算成本较高，这在低资源环境中的应用受到限制。此外，模型在情感细粒度区分和复杂场景的适应性方面也有较大提升空间。未来的研究方向将包括：优化多模态特征融合机制、引入自监督学习以减少对标注数据的依赖、设计轻量化模型以支持实时推理，并深化情感分类的细粒度研究。这些改进将进一步提高模型的性能和跨领域适应能力，推动其在真实场景中的广泛应用。

参考文献

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[2] Mohammed Abbas Ansari, Chandni Saxena, Tanvir Ahmad, et al. Jmi at semeval 2024 task 3: Two-step approach for multimodal ecac using in-context learning with gpt and instruction-tuned llama models. *arXiv preprint arXiv:2403.04798*, 2024.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] Min Cao and Zhendong Wan. Psychological counseling and character analysis algorithm based on image emotion. *IEEE Access*, 2020.

[5] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Jingdong Sun, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *arXiv preprint arXiv:2406.11161*, 2024.

[6] Zhi-Qi Cheng, Yang Liu, Xiao Wu, and Xian-Sheng Hua. Video ecommerce: Towards online video advertising. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1365–1374, 2016.

[7] Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video2shop: Exact matching clothes in videos to online shopping images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4048–4056, 2017.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

[9] Weiquan Fan, Xiangmin Xu, Xiaofen Xing, Weidong Chen, and Dongyan Huang. Lssed: a large-scale dataset and benchmark for speech emotion recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 641–645. IEEE, 2021.

[10] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

[12] Ashley Hutchison and Larry Gerstein. Emotion recognition, emotion expression, and cultural display rules: Implications for counseling. *Journal of Asia Pacific Counseling*, 7(1), 2017.

[13] Maryam Imani and Gholam Ali Montazer. A survey of emotion recognition methods with emphasis on e-learning environments. *Journal of Network and Computer Applications*, 147:102423, 2019.

[14] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[15] Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. Dfew: A large-scale database for recognizing dynamic facial expressions in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2881–2889, 2020.

[16] Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*, 2023.

[17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[18] Zheng Lian, Haiyang Sun, Licai Sun, Jiangyan Yi, Bin Liu, and Jianhua Tao. Affectgpt: Dataset and framework for explainable multimodal emotion recognition. *arXiv preprint arXiv:2407.07653*, 2024.

[19] Zheng Lian, Licai Sun, Yong Ren, Hao Gu, Haiyang Sun, Lan Chen, Bin Liu, and Jianhua Tao. Merbench: A unified evaluation benchmark for multimodal emotion recognition. *arXiv preprint arXiv:2401.03429*, 2024.

[20] Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Shun Chen, Bin Liu, and Jianhua Tao. Gpt-4v with emotion: a zero-shot benchmark for multimodal emotion understanding. *arXiv preprint arXiv:2312.04293*, 2023.

[21] Zheng Lian, Licai Sun, Mingyu Xu, Haiyang Sun, Ke Xu, Zhuofan Wen, Shun Chen, Bin Liu, and Jianhua Tao. Explainable multimodal emotion reasoning. *arXiv preprint arXiv:2306.15401*, 2023.

[22] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[26] Hongxia Xie, Chu-Jun Peng, Yu-Wen Tseng, Hung-Jen Chen, Chan-Feng Hsu, Hong-Han Shuai, and Wen-Huang Cheng. Emovit: Revolutionizing emotion insights with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26596–26605, 2024.

[27] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.