arXiv:2603.01816v1 [cs.MM] 2 Mar 2026

声音、面孔与情感：面向心理健康理解的多模态情绪–认知描述

周志远[1],郭艳荣[1]*，郝世杰[1]*

[1]合肥工业大学 zhouzhiyuan.hfut@gmail.com,
yrguo@hfut.edu.cn, hfut.hsj@gmail.com

摘要

情绪和认知因素对于理解心理健康障碍至关重要。然而，现有方法通常将多模态数据视为分类任务，限制了其可解释性，尤其是在情绪和认知方面。尽管大语言模型（LLM）为心理健康分析带来了机遇，但它们主要依赖文本语义，忽略了多模态输入中细粒度的情绪和认知线索。虽然一些研究通过迁移学习引入情绪特征，但这些特征与心理健康状况之间的联系仍不明确。为解决这些问题，我们提出 ECMC 这一新任务，旨在根据多模态数据生成情绪和认知状态的自然语言描述，并构建情绪–认知画像，从而提升心理健康评估的准确率和可解释性。我们采用编码器—解码器架构，由特定模态的编码器提取特征，再通过基于 Q-former 的双流 BridgeNet 融合这些特征。对比学习增强了情绪和认知特征的提取。随后，LLaMA 解码器将这些特征与标注描述对齐，生成细致的描述文本。大量客观和主观评估表明：1）在生成情绪–认知描述方面，ECMC 优于现有的多模态大语言模型和心理健康模型；2）生成的情绪–认知画像显著提升了心理健康分析中的辅助诊断能力和可解释性。

引言

抑郁症等精神障碍会严重影响身心健康。根据世界卫生组织（WHO）的数据，全球有超过3亿人受到抑郁症影响；到2030年，未经治疗的精神障碍将占疾病总负担的13%（WHO 2017）。许多患者难以获得评估，一个原因是缺乏有效的辅助性心理健康分析方法。目前的诊断方法严重依赖自我报告和临床访谈（Habtamu 等，2023），这类方法既主观又低效。精神科医生短缺，无法满足不断增加的患者需求（Johnson 2022），这凸显了开发高效、可扩展的辅助性心理健康分析方法的迫切需求。

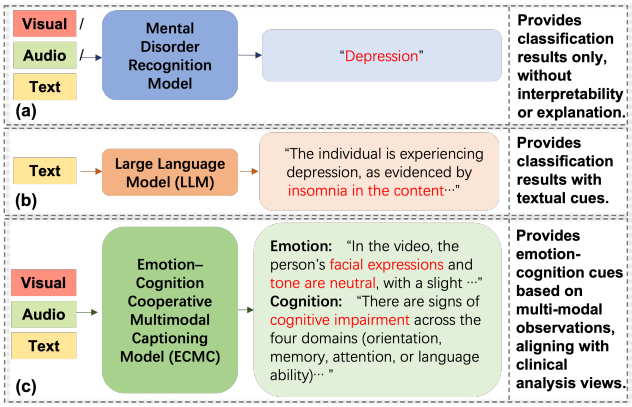


图1：精神障碍检测的不同范式。（a）基于分类。（b）基于大语言模型。（c）本文方法。

然而，现有的辅助性心理健康分析方法面临诸多挑战。首先，大多数方法（Liu 等，2024；Ji 等，2021；Zhang 等，2022）将任务视为分类问题，通过对多模态数据进行分类来判定精神障碍类别，因此缺乏可解释性（图1(a)）。这些方法无法揭示哪些线索与心理健康状况相关。因此，这些方法往往难以令人信服地用于临床诊断支持。其次，大语言模型（LLM）具有很强的自然语言理解能力（Guo 等，2025；Achiam 等，2023）。然而，将其用于心理健康分析时，往往依赖文本信息，例如检测与症状相关的词语（Yang 等，2024；Xu 等，2024a）（图1(b)）。它们无法捕捉来自其他模态的情绪和认知信号，例如面部表情或语音语调，而这些信号在现实临床观察中十分重要。尽管上述两类方法中也有研究（Wu、Zhang 和 Woodland，2023；Wu、Wu 和 Yu，2022）通过迁移学习引入情感特征，但高维特征难以深入揭示情绪与精神障碍之间的关联。这些方法还忽视了认知因素在心理健康分析中的作用。

情绪-认知模式在理解精神障碍方面发挥着至关重要的作用。神经科学研究（Shackman、Fox 和 Seminowicz，2015）表明，精神

*通讯作者。版权所有 © 2026，人工智能促进协会（www.aaai.org）。保留所有权利。

障碍不仅体现在可观察的症状上，也体现在情绪与认知加工的动态机制中。例如，抑郁症通常与长期的情绪停滞和受抑制的认知功能相关，而焦虑障碍则表现出情绪和认知的快速波动，通常表现为惊恐发作。既有研究（Uban、Chulvi 和 Rosso，2021；Li 等，2022）还基于语言特征发现，患有和未患有心理健康状况的个体在情绪和认知特征方面存在显著差异。然而，现有方法很少从多模态视角考察这些模式，也往往无法为临床诊断提供可解释的见解。

为解决这些局限，我们提出情绪–认知协同多模态描述生成（ECMC）任务，旨在生成可解释的情绪-认知档案，以支持心理健康理解（图 1(c)）。受近期情绪描述生成任务研究的启发（Xu 等，2024b；Ye 等，2024），ECMC 要求模型从多种模态中提取与情绪和认知相关的特征，并生成连贯、高质量的自然语言描述。与将复杂信号压缩为类别标签的传统方法不同，ECMC 利用自然语言将多模态信息整合为丰富且具有情境信息的叙述，从而同时呈现情绪状态和认知障碍。这不仅提高了临床医生的可解释性，也支持对不同个体及不同时间进行细粒度比较，使 ECMC 成为实现准确且可解释的心理健康评估的有前景方向。

为此，我们设计了一种多模态编码器–解码器框架，由模态专属编码器、双流 BridgeNet 和基于 LLaMA 的解码器组成。鉴于数据有限，从头训练编码器面临挑战，我们使用针对各模态定制的预训练模型初始化模态专属编码器，以提取初始特征。然而，这类预训练模型通常侧重于帧级表示，无法捕捉与情绪和认知相关的语义。受 BLIP-2（Li 等，2023）启发，我们设计了一种基于 Q-former 的双流 BridgeNet，通过对比学习压缩多模态特征，并解耦情绪与认知表示。最后，我们采用基于 LLaMA 的解码器，利用大语言模型强大的语言理解能力，根据编码后的特征生成情绪–认知描述。训练过程中，解码器引导 BridgeNet 将情绪-认知特征与目标文本空间对齐。

我们的工作贡献总结如下：

- 我们提出了一项新颖的情绪–认知协同多模态描述生成（ECMC）任务，使大语言模型能够生成用于理解心理健康的情绪-认知档案。据我们所知，此视角尚未在以往工作中得到探索。
- 我们为 ECMC[1] 任务设计了一种多模态编码器–解码器框架，可提取情绪和认知特征，并将其与大语言模型的输入空间对齐，从而生成信息丰富的描述文本。
- 大量客观和主观评估表明，我们的方法能够根据多模态观察生成高质量的情绪–认知描述文本。

[1]代码及更多结果：https://github.com/zhouzyhfut/ECMC。

相关工作

精神障碍检测

大多数精神障碍检测方法侧重于设计各种深度学习模型，将样本分类到障碍类别中（Liu 等，2024；Ji 等，2021；Zhang 等，2022）。然而，这些预测背后的根本原因往往是隐含的，限制了其在临床应用中的效用。大语言模型的出现通过在相关数据上进行微调，增强了精神障碍检测（Yang 等，2024；Xu 等，2024a）。尽管如此，这些方法通常仅依赖文本语义信息，无法捕捉患者的面部表情或其他可观察反应。尽管先前的工作（Uban、Chulvi 和 Rosso，2021；Li 等，2022）已通过迁移学习纳入情感信息，但与情绪相关的特征与预测结果之间的关系仍然难以解释。相比之下，我们的方法能够对患者的情绪-认知模式进行多模态观察，并生成显著线索的描述，从而提供更具可解释性的证据，以支持有效的临床诊断。

多模态大语言模型

近年来，多模态大语言模型（MLLMs）在跨模态理解与生成任务中取得了显著进展。例如，BLIP-2（Li 等，2023）引入了一种轻量级的基于查询的 Transformer，有效连接冻结的视觉编码器与语言模型，从而实现出色的零样本表现。Qwen2.5-Omni（Xu 等，2025）支持以视觉内容为基础的流畅多轮对话，彰显了其在多模态场景中的优势。InternVL-2.5（Chen et al.2024）通过密集对齐统一视觉与文本表征，增强了复杂的跨模态推理能力和细粒度理解能力。Sa2VA（Yuan 等，2025）专注于细粒度语义对齐，从而提升了视觉问答和详细图像描述生成的效果。总体而言，多模态大语言模型展现出在多种模态上的出色泛化能力，为下游多模态应用奠定了基础。

自动描述生成

自动描述生成方法通常旨在利用编码器—解码器架构，从视觉或多模态输入中生成自然语言描述。例如，生成情绪—原因感知视频描述有助于捕捉细粒度的多模态线索，进而用于概述对话中的情绪成因（Wang 等，2024）。SECAP（Xu 等，2024b）使用 HuBERT 和 Q-Former 提取聚焦于情绪的语音特征，并利用 LLaMA 生成描述语音情绪的自然语言文本。通过提供细粒度且语义丰富的

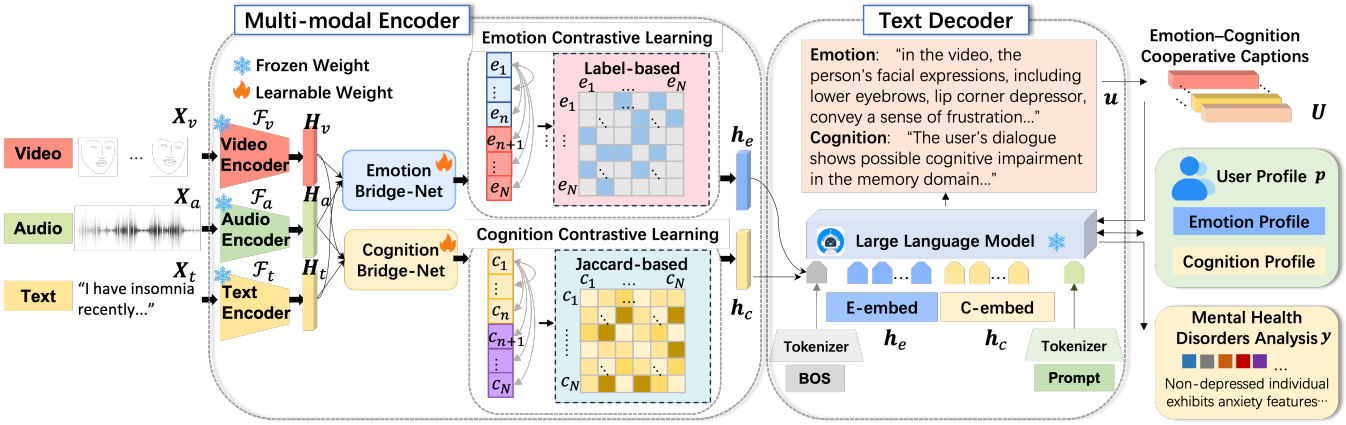


图 2：所提出方法的框架。

文本描述，描述生成任务天然能够提供详细线索，这些线索对于理解心理健康状况极具价值，并有助于深入解读患者的情绪和认知状态。

方法

我们首先概述 ECMC。接着详细说明情绪桥接网络和认知桥接网络如何分别提取情绪特征和认知特征。最后介绍总体训练流程。

模型架构

如图 2 所示，受情绪识别中的自动描述生成启发（Wang 等，2024；Xu 等，2024b；Ye 等，2024），我们采用编码器—解码器架构。

编码器对于给定的话语样本 $\boldsymbol{x}_i = \{\boldsymbol{X}_v, \boldsymbol{X}_a, \boldsymbol{X}_t\}$，其中下标 v、a 和 t 分别表示视频、音频和文本模态，我们首先使用各模态专属的预训练模型提取模态专属的初始表示：

$$\boldsymbol{H}_v = \mathcal{F}_v(\boldsymbol{X}_v), \quad \boldsymbol{H}_a = \mathcal{F}_a(\boldsymbol{X}_a), \quad \boldsymbol{H}_t = \mathcal{F}_t(\boldsymbol{X}_t). \quad (1)$$

在此，我们分别采用 VideoMAE（Tong 等，2022）、HuBERT（van Niekerk 等，2022）和 BERT（Devlin 等，2019年）作为 $\mathcal{F}_v$、$\mathcal{F}_a$ 和 $\mathcal{F}_t$。$\boldsymbol{H}_m \in \mathbb{R}^{T_m \times d_m}, m \in \{v, a, t\}$ 表示各模态的初始表示，其中 T_m 为时间步数量，d_m 为表示的维度。

随后，我们设计了一个双流 BridgeNet，由情绪桥接网络和认知桥接网络组成，用于融合多模态信息并进一步压缩初始表示，因为序列表示往往存在冗余。该过程产生两个嵌入，即 E嵌入和 C嵌入，记为：

$$\{\boldsymbol{h}_e, \boldsymbol{h}_c\} = \mathcal{F}_{bridge}(\boldsymbol{H}_v, \boldsymbol{H}_a, \boldsymbol{H}_t) \quad (2)$$

此处，$\boldsymbol{h}_e \in \mathbb{R}^{1 \times d_e}$ 捕捉与各种情绪类别相关的特征，而 $\boldsymbol{h}_c \in \mathbb{R}^{1 \times d_c}$ 则编码与潜在认知障碍相关的特征。

解码器 随后，我们采用 LLaMA（Touvron 等，2023）作为文本解码器，生成细粒度的情绪—认知描述。借助大语言模型强大的语言理解与生成能力，我们引导 BridgeNet 生成语义丰富的文本特征空间，并产出高质量的描述文本。具体而言，我们在输入前添加一个<BOS>标记，并将其与 E 嵌入 $\boldsymbol{h}_e$ 和 C嵌入 $\boldsymbol{h}_c$拼接，然后接上经过分词的提示。因此，大语言模型的输入可表示为：

$$\boldsymbol{u} = \mathcal{F}_{llm}\big(concat(\texttt{<BOS>}, \boldsymbol{h}_e, \boldsymbol{h}_c, Prompt)\big) \quad (3)$$

从而约束大语言模型的输出空间，以生成准确的描述文本 $\boldsymbol{u}$，使其同时反映话语的情绪与认知方面。

在获得多个话语的情绪—认知描述 $\boldsymbol{U} = \{\boldsymbol{u}_1, ..., \boldsymbol{u}_n\}$ 后，我们指示大语言模型生成用户档案 $\boldsymbol{p}$。该档案总结话语中观察到的显著模式，例如负面情绪和认知障碍：

$$\boldsymbol{p} = \mathcal{F}_{llm}(\boldsymbol{U}, Prompt) \quad (4)$$

用户档案 $\boldsymbol{p}$ 为精神障碍诊断提供了有价值的证据，并进一步帮助通用大语言模型提升其精神障碍检测能力：

$$\boldsymbol{y} = \mathcal{F}_{llm}(\boldsymbol{p}, Prompt, \boldsymbol{X}_t) \quad (5)$$

其中 $\boldsymbol{y}$ 为精神障碍的预测结果。

情感表示提取

为融合多模态特征并进一步压缩初始表示，以提取供解码器使用的情感表示，我们设计了 Emotion BridgeNet，如图3所示。

Q-former 对于每种模态 m，我们引入一组可学习的查询令牌 $\boldsymbol{Q}_m \in \mathbb{R}^{l_q \times d_q}$，其中 l_q 是查询令牌的数量，d_q 是隐藏维度。我们首先对查询令牌应用自注意力，以建模查询内部的依赖关系：

$$\boldsymbol{Q}'_m = \text{softmax}\left(\frac{\boldsymbol{Q}_m \boldsymbol{W}_q (\boldsymbol{Q}_m \boldsymbol{W}_k)^\top}{\sqrt{d_k}}\right) \boldsymbol{Q}_m \boldsymbol{W}_v \quad (6)$$

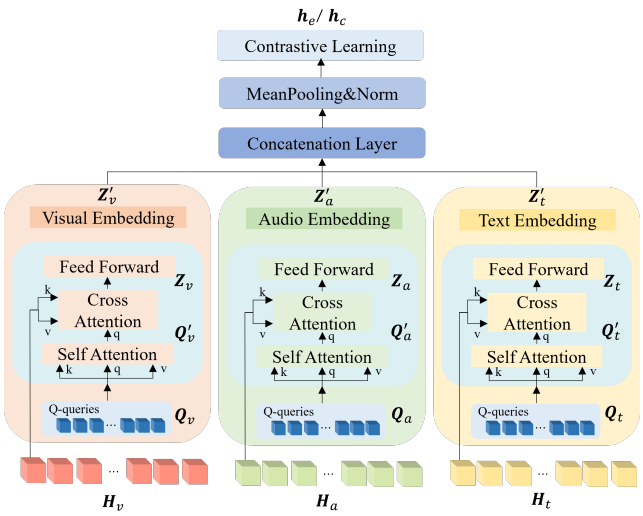


图3：BridgeNet 的架构。

其中 $\boldsymbol{W}_q \in \mathbb{R}^{d_q \times d_k}$，$\boldsymbol{W}_k \in \mathbb{R}^{d_q \times d_k}$，$\boldsymbol{W}_v \in \mathbb{R}^{d_q \times d_v}$分别是查询、键和值对应的可学习权重矩阵。接下来，将更新后的查询 $\boldsymbol{Q}'_m \in \mathbb{R}^{l_q \times d_v}$ 用作交叉注意力的查询：

$$\boldsymbol{Z}_m = \text{softmax}\left(\frac{\boldsymbol{Q}'_m \boldsymbol{W}_q'^{(c)} (\boldsymbol{H}'_m \boldsymbol{W}_k'^{(c)})^\top}{\sqrt{d_k}}\right) \boldsymbol{H}'_m \boldsymbol{W}_v'^{(c)} \tag{7}$$

其中，$\boldsymbol{H}'_m = \boldsymbol{H}_m \boldsymbol{W}_m$、$\boldsymbol{W}'_{(q^c)} \in \mathbb{R}^{d_v \times d_k}$）、$\boldsymbol{W}'_{(c)\ k} \in \mathbb{R}^{d_q \times d_k}$、$\boldsymbol{W}'_{(v^c)} \in \mathbb{R}^{d_q \times d_v}$）、$\boldsymbol{W}_m \in \mathbb{R}^{d_m \times d_q}$是可学习的权重矩阵。随后，将经过注意力处理的查询表示$\boldsymbol{Z}_m \in \mathbb{R}^{l_q \times d_v}$ 输入前馈层：

$$\boldsymbol{Z}'_m = ReLU(\boldsymbol{Z}_m \boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{8}$$

其中，$\boldsymbol{W}_1 \in \mathbb{R}^{d_v \times d_e}$、$\boldsymbol{W}_2 \in \mathbb{R}^{d_e \times \frac{d_e}{3}}$和 $\boldsymbol{b}_1$，$\boldsymbol{b}_2$ 是偏置。

我们将这些表示 $\boldsymbol{Z}'_m$ 拼接起来，并将其投影到特征空间：

$$\boldsymbol{h}_e = Norm(Pooling(Concat(\boldsymbol{Z}'_v, \boldsymbol{Z}'_a, \boldsymbol{Z}'_t)\boldsymbol{W}_3 + \boldsymbol{b}_3)), \tag{9}$$

其中，$\boldsymbol{h}_e \in \mathbb{R}^{l_d \times d_e}$、$\boldsymbol{W}_3 \in \mathbb{R}^{d_e \times d_e}$ 是投影矩阵，$\boldsymbol{b}_3$ 是偏置。Pooling 表示均值池化操作，Norm 表示 L2归一化。

情绪对比学习对于精神障碍分析，心理学研究（Joormann and Stanton 2016；Vanderlind et al. 2020；Zakiei et al. 2014）指出，负面情绪是一个显著特征。因此，我们根据情感空间中的效价维度（Kensinger 2004），将特征表示 $\boldsymbol{h}_{e,\ i}$ 划分为负面、中性和正面类别。为了更好地捕捉不同情绪类别的表示，我们引入一种基于标签匹配的对比学习方法。首先，我们计算相似度矩阵：$\boldsymbol{S}_{ij} = \boldsymbol{h}_e^\top{}_{,\ i} \boldsymbol{h}_{e,\ \tau^{\ j}}$. 然后构建一个二值掩码 $\boldsymbol{M}_{ij} = 1$，当 $y_{e,\ i} = y_{e,\ j}$ 且 $i \neq j$ 时取值为 1，否则为 0，其中 $y_e \in \{-1, 0, 1\}$ 表示情绪类别标签，用于识别类内正样本。损失结合了类内紧凑性和类间可分性：

$$\begin{aligned}\mathcal{L}_{\text{emo}} = &-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\{j : \boldsymbol{M}_{ij} = 1\}|}\sum_{j:\boldsymbol{M}_{ij}=1}\log p_{ij} \\ &+\frac{1}{N}\sum_{i=1}^{N}\log\Big(1 + \sum_{j:\boldsymbol{M}_{ij}=0}\exp(\boldsymbol{S}_{ij})\Big),\end{aligned} \tag{10}$$

其中，$\log\ p_{ij} = \boldsymbol{S}_{ij} - \log \sum_{k \neq i} \exp(\boldsymbol{S}_{ik})$，$N$表示批次大小。

认知表征提取

与情感表示提取过程类似，我们采用 Q-former 对多模态初始特征 $\boldsymbol{H}_m$ 进行压缩和融合，得到认知表征 $\boldsymbol{h}_c$。在临床认知量表简易精神状态检查（MMSE）（Arevalo-Rodriguez 等，2015）的指导下，我们聚焦于四种特定的认知障碍：定向力缺损、注意力缺陷、记忆力缺损和语言障碍，并旨在发现与这些不同认知类别相对应的表征。由于一个样本$\boldsymbol{h}_{c,\ i}$ 可能同时关联多种认知障碍，我们设计了一种基于 Jaccard 相似度的多标签对比学习目标，以更好地捕捉这些障碍的共性与差异。对于认知特征 $\boldsymbol{h}_{c,\ i}$，每个样本可能有多个标签 $\boldsymbol{y}_{c,\ i} \in \{0, 1\}^4$。我们用基于 Jaccard 指数的软相似度权重替代二元掩码：

$$\boldsymbol{W}_{ij} = \begin{cases} 1, & \boldsymbol{y}_{c,i} = \boldsymbol{y}_{c,j} = \varnothing \\ 0, & \boldsymbol{y}_{c,i} = \varnothing \text{ or } \boldsymbol{y}_{c,j} = \varnothing \\ \dfrac{|\boldsymbol{y}_{c,i} \cap \boldsymbol{y}_{c,j}|}{|\boldsymbol{y}_{c,i} \cup \boldsymbol{y}_{c,j}|}, & \text{otherwise.} \end{cases} \tag{11}$$

该损失函数会柔性地拉近标签重叠度较高的样本，并推远其他样本：

$$\begin{aligned}\mathcal{L}_{\text{cog}} = &-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\sum_j \exp(\boldsymbol{S}_{ij}) \cdot \boldsymbol{W}_{ij}}{\sum_j \exp(\boldsymbol{S}_{ij})} \\ &+\frac{1}{N}\sum_{i=1}^{N}\log\Big(1 + \sum_j \exp(\boldsymbol{S}_{ij})(1 - \boldsymbol{W}_{ij})\Big).\end{aligned} \tag{12}$$

训练过程

为利用大语言模型增强多模态情感–认知描述，我们设计了一个两阶段训练过程。在第一阶段，我们联合训练情感表示提取和认知表征提取过程。这种协同训练能够有效压缩并融合多模态信息，从而提取紧凑的情感和认知表示：

$$\mathcal{L}_1 = \mathcal{L}_{\text{emo}} + \mathcal{L}_{\text{cog}}, \tag{13}$$

其中，我们冻结模态特定编码器的参数，并使用 BERT（Devlin 等，2019年）的预训练参数初始化 Q-Former。

方法	模态	领域	BLEU-1	BLEU-2	BLEU-4	METEOR	ROUGE-L	CIDEr	F BERT
InternVL-2.5-8B	VT	通用	8.35	3.30	0.96	10.06	13.98	0.96	6.48
DeepSeek-VL-2-small	VT	通用	2.98	1.33	0.42	5.99	10.00	0.42	2.02
Sa2VA-8B	VT	通用	14.36	7.05	2.14	15.35	20.34	2.14	12.28
Qwen2.5-Omni-7B	AVT	通用	12.74	4.86	1.42	13.31	16.99	1.42	8.93
PsycoLLM	T	Mental	4.86	1.76	0.50	5.68	7.68	0.50	0.35
MindChat	T	Mental	5.44	1.75	0.50	6.73	9.71	0.50	0.87
EmoLLM	T	Mental	8.50	3.50	1.19	10.08	14.25	1.19	6.25
CPsyCoun	T	Mental	17.44	7.24	2.33	15.07	18.90	2.33	9.30
Ours	AVT		**34.76**	**19.40**	**8.28**	**29.47**	**24.91**	**8.28**	**27.13**

表 1：与其他方法的情感描述性能比较。A、V、T 分别表示音频、视频和文本。粗体表示最佳结果，下划线表示次佳结果。

方法	模态	领域	BLEU-1	BLEU-2	BLEU-4	METEOR	ROUGE-L	CIDEr	F BERT
InternVL-2.5-8B	VT	通用	8.17	3.05	0.84	9.35	16.52	0.84	18.43
DeepSeek-VL-2-small	VT	通用	8.49	4.40	1.72	10.34	15.45	1.72	6.75
Sa2VA-8B	VT	通用	15.82	6.24	1.61	13.15	22.59	1.61	18.78
Qwen2.5-Omni-7B	AVT	通用	12.82	5.30	1.48	12.04	19.56	1.48	20.03
PsycoLLM	T	Mental	6.22	3.19	1.30	10.59	10.48	1.30	6.60
MindChat	T	Mental	6.81	3.02	1.11	8.33	11.74	1.11	5.41
EmoLLM	T	Mental	17.00	6.66	1.98	13.29	19.08	1.98	14.99
CPsyCoun	T	Mental	9.32	3.89	1.06	18.88	23.07	1.06	4.47
Ours	AVT		**35.92**	**25.05**	**15.32**	**35.86**	**39.82**	**15.32**	**41.04**

表2：认知描述性能与其他方法的比较

在第二阶段，我们微调情绪桥接网络和认知桥接网络，以有效地将提取的特征整合到大语言模型中。具体而言，我们冻结模态特定编码器和大语言模型的参数，并采用交叉熵损失，使大语言模型生成带注释的描述文本 $\hat{\boldsymbol{u}}$，从而引导桥接网络生成与大语言模型输入空间对齐的特征：

$$\mathcal{L}_2 = CELoss(\hat{\boldsymbol{u}}, \boldsymbol{u}) \tag{14}$$

数据集与评估指标

数据集 我们在 MMDA 数据集（Jiang、Zhang 和 Sun，2022）上评估了我们的方法。该数据集是一个用于精神障碍分析的大规模多模态基准，包含 1,025 个经临床验证的样本。每个样本均包含来自临床访谈的视觉、声学和文本数据，并依据标准化评估（HAMD 和 HAMA）标注抑郁症和焦虑症。为生成话语级别的情绪—认知描述文本，我们结合了自动标注与人工修订。首先将访谈切分为问答轮次。随后，Emotion-LLaMA（Cheng 等，2024）生成融合面部、声音和语义线索的情绪描述。DeepSeek-671B（Guo 等，2025）利用这些输出和文本内容，生成相应的认知描述文本以及情绪和认知类别标签，最终得到 30,592 对带注释的话语—描述文本。为确保数据质量，我们移除了非信息性片段（例如问候、询问个人信息、闲谈或时长不足 2 秒的话语），以保留核心医患互动。标注人员还进一步剔除了不一致的标签，以尽可能减少噪声和偏差。我们计算了评审者间一致性，结果显示一致性较高（情绪的 Fleiss's $\kappa = 0.64$，认知的 0.73）。经过筛选和验证后，数据集包含 13,536 个训练片段、1,402 个验证片段和 3,790 个测试片段。数据集的类别分布为：定向力缺损占 0.91%，注意力缺损占 8.54%，记忆力缺损占 10.38%，语言障碍占 15.89%，负面情绪占 31.85%。

客观指标 对于 ECMC 任务，我们采用自动描述生成中使用的自然语言评估指标，包括 BLEU-1、BLEU-2、BLEU-4、METEOR、ROUGE-L、CIDEr 和 F BERT。其中，BLEU、METEOR、ROUGE-L 和 CIDEr 主要是词级匹配指标，而 F BERT 是基于语义相似度的指标。对于生成的情绪-认知画像在心理障碍检测中的评估，我们采用准确率（ACC）和 F1分数（F1）。

主观指标 为评估数据集标注的质量以及不同方法生成的描述文本的有效性，我们采用五点李克特量表（Joshi等人，2015）进行人工评估，评估维度如下：（1）情感准确性（EAcc），衡量生成的情绪描述文本是否准确反映视频内容；（2）认知准确率（CAcc），衡量关于认知障碍的描述文本是否与对话内容相符；（3）细节丰富度（Detail），评估描述文本中细节的丰富程度和粒度；（4）标签一致性（Consis），评估描述文本是否与相应的分类标签一致；以及（5）总体质量

方法	准确率	F1分数	方法	准确率	F1分数	方法	准确率	F1分数	图像准确率	图像F1分数
Llama-3-8B	44.95	46.43	Qwen3-8B	50.46	60.29	GLM4-9B	47.71	58.39	0.00	0.00
InternVL-2.5-8B	48.62	46.15	InternVL-2.5-8B	69.72	69.81	InternVL-2.5-8B	53.21	59.84	9.48↑	3.56↑
DeepSeek-VL-2-small	41.28	47.54	DeepSeek-VL-2-small	51.38	59.54	DeepSeek-VL-2-small	51.38	60.15	0.31↑	0.71↑
Sa2VA-8B	48.62	49.09	$a2VA-8B	67.89	69.03	$a2VA-8B	57.80	62.90	10.40↑	5.30↑
Qwen2.5-Omni-7B	49.54	54.24	Qwen2.5-Omni-7B	65.14	67.24	Qwen2.5-Omni-7B	56.88	58.46	9.48↑	4.94↑
PsycoLLM	47.71	44.66	PsycoLLM	54.13	62.12	PsycoLLM	53.21	61.07	3.98↑	0.91↑
MindChat	29.36	36.36	MindChat	58.72	62.81	MindChat	57.80	62.90	0.92↑	1.01↓
EmoLLM	45.87	54.96	EmoLLM	61.47	63.16	EmoLLM	55.05	63.16	6.42↑	5.39↑
CPsyCounX	48.62	48.15	CPsyCounX	67.89	68.47	CPsyCounX	55.05	60.80	9.48↑	4.10↑
Ours	**50.46**	**56.69**	**我们的方法**	**70.64**	**70.27**	**我们的方法**	**59.63**	**65.62**	**12.54↑**	**9.16↑**

表3：基线大语言模型使用不同方法所得情绪-认知档案进行抑郁症检测。第1行：基线模型的结果。'图像准确率'和'图像F1分数'显示相对于Llama-3-8B、Qwen3-8B和GLM4-9B在准确率和F1分数上的平均增益。

方法	准确率	F1分数	方法	准确率	F1分数	方法	准确率	F1分数	图像准确率	图像F1分数
Llama-3-8B	38.54	52.80	Qwen3-8B	41.67	57.58	GLM4-9B	43.75	57.14	0.00	0.00
InternVL-2.5-8B	56.25	56.25	InternVL-2.5-8B	55.21	63.87	InternVL-2.5-8B	42.71	58.02	10.07↑	3.54↑
DeepSeek-VL-2-small	48.96	55.05	DeepSeek-VL-2-small	45.83	59.38	DeepSeek-VL-2-small	41.67	57.58	4.17↑	1.50↑
Sa2VA-8B	55.21	52.75	$a2VA-8B	54.17	62.71	$a2VA-8B	44.79	58.46	10.07↑	2.13↑
Qwen2.5-Omni-7B	50.00	50.00	Qwen2.5-Omni-7B	65.62	65.96	Qwen2.5-Omni-7B	43.75	58.46	11.80↑	2.30↑
PsycoLLM	52.08	57.41	PsycoLLM	48.96	59.50	PsycoLLM	43.75	58.46	6.94↑	2.62↑
MindChat	47.92	46.81	MindChat	46.88	59.84	MindChat	41.67	57.58	4.17↑	1.10↓
EmoLLM	54.17	55.10	EmoLLM	61.46	66.06	EmoLLM	43.75	58.46	11.81↑	4.03↑
CPsyCounX	46.88	54.05	CPsyCounX	56.25	63.16	CPsyCounX	42.71	58.02	7.29↑	2.57↑
Ours	**57.29**	**58.59**	**我们的方法**	**66.67**	**69.16**	**我们的方法**	**44.95**	**58.91**	**14.98↑**	**6.38↑**

表 4：使用不同方法生成的情绪-认知档案进行焦虑检测的基线大语言模型

（Overall），对质量进行总体评估。

实验与分析

实验设置

第一阶段，我们在两张 NVIDIA RTX A6000 GPU 上训练模型，批次大小为 64，最多训练 500 个轮次。音频的最大序列长度为 1024，视频的最大序列长度为 512。我们使用 AdamW 优化器，学习率为 1.3×10^{-5}，权重衰减为 1.0×10^{-6}。模型总参数量约为 7.01 亿，其中约 5.98 亿个参数可训练。第二阶段，我们训练模型时，每张 GPU 的微批次大小为 8，并累积 8 个步骤。我们还采用 DeepSpeed ZeRO 第 2 阶段优化，并设置混合精度训练（fp16）。在此阶段，模型总参数量约为 76 亿，其中约 6.05 亿个参数可训练。

性能分析

我们通过与基线方法进行描述比较、评估精神障碍检测性能、对模态和对比模块进行消融实验、分析不同障碍的统计数据，以及开展人工评估来评估我们的方法。

情绪–认知描述对比 我们选择了通用多模态大语言模型（InternVL-2.5（Chen 等，2024）、DeepSeek-VL-2（Lu 等，2024）、Sa2VA（Yuan 等，2025）、Qwen2.5-Omni（Xu 等，2025）），以及专门针对心理健康领域进行微调的大语言模型（PsycoLLM（Hu 等，2024）、MindChat[2]，EmoLLM[3],CPsyCoun（Zhang 等，2024））进行比较。表 1 和表 2 展示了不同模型在 ECMC 上的表现。可以看出，我们的模型在情绪描述和认知描述方面均始终优于其他竞争模型。Sa2VA 和 Qwen2.5-Omni 等通用领域多模态大语言模型得益于更强的指令遵循能力和更广泛的数据覆盖，表现相对较好，但往往生成简短或不一致的输出。心理健康大语言模型虽然针对特定领域，但仅限于文本，且难以描述非语言线索，因此得分较低。相比之下，我们的方法能有效整合面部表情、语音语调和文本语义等多模态数据，从而生成更全面、质量更高的情绪—认知描述。

对精神障碍检测的影响 我们使用现有的基线大语言模型评估不同方法生成的情绪-认知档案对抑郁症和焦虑症检测的影响，包括 LLaMA-3-8B（Dubey 等，2024）、Qwen3-8B（Yang 等，2025）和 GLM4-9B（Zeng 等，2024）。对于每位受试者，我们首先使用 DeepSeek-671B 将不同方法生成的描述文本（表 1 和表 2）总结为情绪-认知档案。随后，将这些档案与原始对话内容一同作为输入提供给基线大语言模型。

[2]https://github.com/X-D-Lab/MindChat [3] https://github.com/SmartFlowAI/EmoLLM

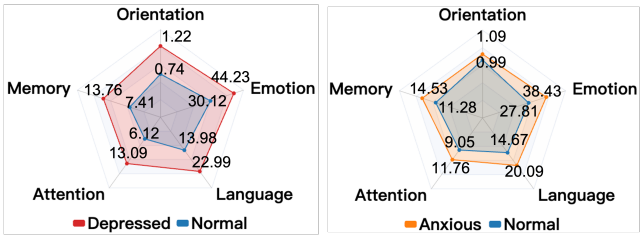


图4：正常个体与不同精神障碍患者之间的情绪和认知差异

音频	视频	Text	情绪对比学习	认知对比学习	F_BERT
✓			✓	✓	18.10
	✓		✓	✓	16.30
		✓	✓	✓	16.75
✓	✓		✓	✓	22.05
✓		✓	✓	✓	29.30
	✓	✓	✓	✓	27.21
✓	✓	✓			23.89
✓	✓	✓		✓	27.24
✓	✓	✓	✓		26.94
✓	✓	✓	✓	✓	34.09

表5：不同模块的消融实验。EmoCL 和 CogCL 分别表示情绪对比学习和认知对比学习。

如表 3 和表 4 所示，我们的方法生成的档案在两种障碍类型上的准确率和 F1分数均更高，显著提升了不同基线方法的性能：在抑郁症检测中，平均准确率和 F1分数分别提升了 12.54% 和 9.16%；在焦虑检测中，分别提升了 14.98% 和 6.38%。值得注意的是，我们还观察到，无效的描述文本甚至会降低检测性能，因为冗长但缺乏信息的文本可能会增加大语言模型提取相关指标的难度。结果验证了我们的方法能够生成质量更高的情绪-认知档案，从而有效提升使用现有通用模型进行精神障碍检测的准确率。

消融实验 我们开展消融实验，以验证不同模态和对比学习模块对最终性能的贡献。我们采用语义相似度指标 F BERT，评估情绪–认知描述的总体效果。如表5所示，在单模态输入中，音频表现最佳，说明其具有较强的判别能力。多模态组合进一步提升了结果，其中音频+文本展现出很强的互补性。移除 EmoCL 或 CogCL 中任一模块，都会使模型退化为单分支结构，导致性能下降；同时移除两者则会造成显著退化。这表明，双分支结构对于有效解耦情绪与认知特征至关重要。采用全部模态和对比学习目标的完整模型得分最高，证实了各组件的有效性。

方法/数据集	EAcc	CAcc	细节	一致性	总体
数据集	4.2	3.9	4.6	4.3	4.3
Sa2VA	3.76	4.40	**4.36**	-	4.15
Qwen2.5-Omni	3.73	4.45	4.32	-	4.12
EmoLLM	3.6	3.91	3.96	-	3.79
CPsyCoun	3.2	3.6	3.68	-	3.46
Ours	**3.98**	**4.51**	4.32	-	**4.26**

表6：数据集与方法的人工评估

ods

不同精神障碍的情绪与认知统计 为探究不同精神障碍患者与健康个体在情绪—认知模式上的差异，我们计算了每位受试者话语中负面情绪、定向力缺损、注意力缺陷、记忆力缺损和语言障碍所占的比例，再分别计算各组（抑郁症、焦虑症和健康对照组）的平均值。如图4所示，与健康参与者相比，抑郁症和焦虑症患者的负面情绪比例均显著更高。在认知障碍方面，抑郁症患者相关表现的出现频率明显高于焦虑症患者。在各类认知障碍中，定向力缺损出现的频率较低，而注意力缺陷、记忆力缺损和语言障碍更为常见。以上观察结果验证了 ECMC 任务的必要性。

人工评估 我们邀请了八位具有心理健康领域专长的心理学从业者，对我们的数据集和模型输出进行评估。如表6所示，数据集的总体得分为4.3，表明情绪与认知描述文本质量较高。EAcc、细节和一致性得分均高于4.0，说明情绪内容准确、细节丰富，且标签可靠，适用于对比学习。CAcc得分略低，反映出从多模态和文本输入中生成认知准确描述具有一定难度，也凸显了改进多模态认知建模的必要性。我们将本方法与两种表现优异的通用多模态大语言模型和两种领先的心理健康大语言模型进行了比较。本模型在EAcc、CAcc和总体得分上均为最高，在细节方面排名第二。值得注意的是，多模态大语言模型获得的主观评分高于心理健康大语言模型，这从经验层面验证了结合视觉和声学线索的有效性，而心理健康大语言模型缺乏这些线索。

结论

本文提出ECMC任务，旨在从多模态数据生成可解释的情绪-认知档案，以服务于心理健康理解。为支持该任务，我们设计了一种新颖的编码器-解码器框架，通过基于LLaMA的解码器提取并对齐情绪-认知特征。实验结果表明，我们的方法能够生成高质量、可解释的描述文本，为可解释且细粒度的心理健康评估提供了一个有前景的方向。

致谢

本工作得到国家重点研发计划（资助编号：2023YFC2506800）以及国家自然科学基金委员会（资助编号：62072152、62172137）的支持。

参考文献

Achiam, J.; Adler,S.; Agarwal, S.;Ahmad,L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; 等. 2023. GPT-4技术报告。*arXiv*预印本 *arXiv:2303.08774*。

Arevalo-Rodriguez, I.; Smailagic, N.; i Figuls, M. R.; Ciap- poni, A.; Sanchez-Perez, E.; Giannakou, A.; Pedraza, O. L.; Cosp, X. B.; and Cullum, S. 2015. 简易精神状态检查（MMSE）用于检测轻度认知障碍（MCI）患者的阿尔茨海默病及其他痴呆。科克伦系统评价数据库，(3)。

Chen, Z.; Wu,J.; Wang, W.; Su,W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; 等. 2024. Internvl：扩展视觉基础模型并使其适用于通用视觉语言任务。在*IEEE/CVF*计算机视觉与模式识别会议论文集，24185–24198。

Cheng, Z.; Cheng, Z.-Q.; He, J.-Y.; Wang, K.; Lin, Y.; Lian, Z.; Peng, X.; and Hauptmann, A. 2024. Emotion-llama：通过指令微调实现多模态情绪识别与推理。神经信息处理系统进展，37：110805–110853。

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert：用于语言理解的深度双向 Transformer 预训练。In*2019*年北美计算语言学协会年会论文集：人类语言技术，第*1*卷（长篇与短篇论文），*4171–4186*。

Dubey, A.; Jauhri, A.;Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. Llama 3 模型家族。*arXiv* 电子预印本，arXiv–2407。

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; 等. 2025年. DeepSeek-R1：通过强化学习激励大语言模型的推理能力。*arXiv* 预印本 *arXiv:2501.12948*。

Habtamu, K.; Birhane, R.; Demissie, M.; 和 Fekadu, A. 2023. 初级卫生保健中改善抑郁症检测的干预措施：系统综述。《*Systematic Reviews*》，12(1)：25。

Hu, J.; Dong, T.; Gang, L.; Ma, H.; Zou, P.; Sun, X.; Guo, D.; Yang, X.; 和 Wang, M. 2024. PsycoLLM：增强大语言模型的心理理解与评估能力。*IEEETransactions onComputationalSocialSystems*。

Ji, S.; Zhang,T.; Ansari, L.; Fu, J.; Tiwari, P.; and Cambria, E. 2021. MentalBERT：面向精神卫生保健的公开可用预训练语言模型。*arXiv* 预印本 *arXiv:2110.15621*。

Jiang, Y.; Zhang, Z.; and Sun, X. 2022. MMDA：用于抑郁症和焦虑症检测的多模态数据集。载于国际模式识别会议论文集，691–702。施普林格。Johnson, R. M. 2022. 应对精神科人力短缺问题。《*Psychiatric Times*》。

Joormann, J.; and Stanton, C. H. 2016. 抑郁症中的情绪调节研究：综述与未来方向。《行为研究与治疗》，86：35–49。Joshi, A.; Kale, S.; Chandel, S.; and Pal, D. K. 2015. 李克特量表：探析与阐释。《英国应用科学与技术杂志》，7(4)：396。

Kensinger, E. A. 2004. 记忆情绪体验：效价和唤醒的作用。《神经科学评论》，15(4)：241–252。

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2：使用冻结图像编码器和大语言模型引导语言-图像预训练。载于论文集 *of*

国际机器学习会议，19730– 19742。PMLR。

Li, M.; Zhang, J.; Song, J.; Li, Z.; and Lu, S. 2022. 一种基于情绪冲突认知行为的临床导向型非重度抑郁诊断方法。*IEEE*计算社会系统汇刊，10(1)：131– 141。

Liu, H.; Li, C.; Zhang, X.; Zhang, F.; Wang, W.; Ma, F.; Chen, H.; Yu, H.; and Zhang, X. 2024. 基于胶囊网络与对比学习的抑郁症检测。载于*AAAI*人工智能会议论文集，22231–22239。

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; 等. 2024. Deepseek-vl：迈向真实世界的视觉语言理解。*arXiv* 预印本 *arXiv:2403.05525*。

Shackman, A. J.; Fox, A. S.; 和 Seminowicz, D. A. 2015. 认知情绪脑：理解神经精神障碍的机遇与挑战。行为与脑科学，38：e86。

Tong, Z.; Song, Y.; Wang, J.; 和 Wang, L. 2022. Video-MAE：掩码自编码器是适用于自监督视频预训练的数据高效学习器。神经信息处理系统进展，35：10078–10093。

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozi`ere, B.; Goyal, N.; Hambro, E.; Azhar, F.; 等. 2023. Llama：开放且高效的基础语言模型。*arXiv*预印本 *arXiv:2302.13971*。

Uban, A.-S.; Chulvi, B.; and Rosso, P. 2021. 基于社交媒体数据的心理健康障碍情绪与认知分析。《未来一代计算机系统》，124：480–494。

van Niekerk, B.; Carbonneau, M.-A.; Za¨ıdi, J.; Baas, M.; Seut´e, H.; and Kamper, H. 2022年. 改进语音转换中离散语音单元与软语音单元的比较. 载于*IEEE*声学、语音与信号处理国际会议论文集。

Vanderlind, W. M.; Millgram, Y.; Baskin-Sommers, A. R.; Clark, M. S.; and Joormann, J. 2020年. 理解抑郁症中的积极情绪缺失：从情绪偏好到情绪调节. 《临床心理学评论》, 76: 101826. Wang, F.; Ma, H.; Shen, X.; Yu, J.; and Xia, R. 2024. 先观察，再生成：用于对话中多模态情绪原因生成的情绪原因感知视频描述. 载于第*32*届*ACM*国际多媒体会议论文集, 5820–5828. WHO. 2017年. 抑郁症及其他常见精神障碍：全球卫生估算. 技术报告，世界卫生组织. Wu, W.; Wu, M.; and Yu, K. 2022年. 气候与天气：通过情绪识别考察抑郁症检测. 载于*IEEE*国际声学、语音与信号处理会议论文集, 6262–6266. IEEE. Wu, W.; Zhang, C.; and Woodland, P. C. 2023. 基于语音的抑郁症检测中的自监督表征. 载于*IEEE*声学、语音与信号处理国际会议论文集, 1–5. Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; 等. 2025年. Qwen2. 5-omni技术报告. *arXiv*预印本 *arXiv:2503.20215*. Xu, X.; Yao, B.; Dong, Y.; Gabriel, S.; Yu, H.; Hendler, J.; Ghassemi, M.; Dey, A. K.; and Wang, D. 2024a. Mental-LLM：利用大语言模型通过在线文本数据进行心理健康预测. 载于*ACM*交互、移动、可穿戴与泛在技术会议录. Xu, Y.; Chen, H.; Yu, J.; Huang, Q.; Wu, Z.; Zhang, S.-X.; Li, G.; Luo, Y.; and Gu, R. 2024b. SECAP：利用大语言模型进行语音情绪描述. 载于*AAAI*人工智能会议论文集, 19323–19331. Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; 等. 2025年. Qwen3技术报告. *arXiv*预印本 *arXiv:2505.09388*. Yang, K.; Zhang, T.; Kuang, Z.; Xie, Q.; Huang, J.; and Ananiadou, S. 2024. MentaLLaMA：利用大语言模型对社交媒体进行可解释的心理健康分析. 载于*2024*年*ACM Web*会议论文集, 4489–4500. Ye, C.; Chen, W.; Li, J.; Zhang, L.; and Mao, Z. 2024. 用于情感视频描述的双路径协同生成网络. 载于第*32*届*ACM*国际多媒体会议论文集, 496–505. Yuan, H.; Li, X.; Zhang, T.; Huang, Z.; Xu, S.; Ji, S.; Tong, Y.; Qi, L.; Feng, J.; and Yang, M.-H. 2025年. Sa2VA：将SAM2与LLaVA结合，实现图像与视频的密集视觉定位理解. *arXiv*预印本 *arXiv:2501.04001*. Zakiei, A.; Karami, J.; Alikhani, M.; Ghasemi, S.; and Gilan, N. R. 2014年. 述情障碍、非理性信念、积极与消极情绪与精神障碍之间的关系. 《抑郁与焦虑》, 17: 18. Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Zhang, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; Yu, H.; Wang, H.; Sun, J.; Zhang, J.; Cheng, J.; Gui, J.; Tang, J.; Zhang, J.; Sun, J.; Li, J.; Zhao, L.; Wu, L.; Zhong, L.; Liu, M.; Huang, M.; Zhang, P.; Zheng, Q.; Lu, R.; Duan, S.; Zhang, S.; Cao, S.; Yang, S.; Tam, W. L.; Zhao, W.; Liu, X.; Xia, X.; Zhang, X.; Gu, X.; Lv, X.; Liu, X.; Liu, X.; Yang, X.; Song, X.; Zhang, X.; An, Y.; Xu, Y.; Niu, Y.; Yang, Y.; Li, Y.; Bai, Y.; Dong, Y.; Qi, Z.; Wang, Z.; Yang, Z.; Du, Z.; Hou, Z.; and Wang, Z. 2024. ChatGLM：从 GLM-130B 到 GLM-4 All Tools 的大语言模型系列。arXiv:2406.12793。Zhang, C.; Li, R.; Tan, M.; Yang, M.; Zhu, J.; Yang, D.; Zhao, J.; Ye, G.; Li, C.; and Hu, X. 2024. CPsyCoun：一种面向中文心理咨询的基于报告的多轮对话重建与评估框架。*arXiv* 预印本 *arXiv:2405.16433*。Zhang, Z.; Chen, S.; Wu, M.; and Zhu, K. Q. 2022. 心理量表引导的风险帖子筛查，用于抑郁症早期检测。*arXiv* 预印本 *arXiv:2205.09497*。