arXiv:2602.05496v1 [cs.MM] 5 Feb 2026

XEmoGPT：一种具备线索级感知与推理能力的可解释多模态情绪识别框架

Hanwen Zhang、Yao Liu、P eiyuan Jiang、Lang Junjie、Xie Jun、Yihui He、Yajiao Deng、Siyu Du 和 Qiao Liu

摘要——可解释多模态情感识别在人机交互和社交媒体分析等应用中发挥着至关重要的作用。然而，当前方法在情感线索的线索级感知与推理方面面临困难，主要有两大挑战：1）通用模态编码器的预训练目标是捕捉全局结构和通用语义，而非细粒度情感线索，因此对情感信号的敏感度有限；2）现有数据集通常需要在标注质量与规模之间权衡，导致针对情感线索的监督不足，最终限制了线索级推理。此外，现有评估指标不足以评估线索级推理性能。为应对这些挑战，我们提出可解释情绪 GPT（XEmoGPT），这是一种新型 EMER 框架，能够感知并推理情感线索。该框架包含两个专用模块：视频情感线索桥（VECB）和音频情感线索桥（AECB），通过精心设计的任务增强视频和音频编码器，以实现细粒度情感线索感知。为进一步支持线索级推理，我们构建了大规模数据集 EmoCue，旨在教会 XEmoGPT 如何对多模态情感线索进行推理。此外，我们还推出 EmoCue-360，这是一种通过语义相似度提取并匹配情感线索的自动化指标；并发布 EmoCue-Eval，这是一个包含 400 个专家标注样本的基准，涵盖多样的情感场景。实验结果表明，XEmoGPT 在情感线索感知和推理方面均取得了出色的表现。

索引词—多模态大语言模型、可解释的多模态情绪识别、模型可解释性。

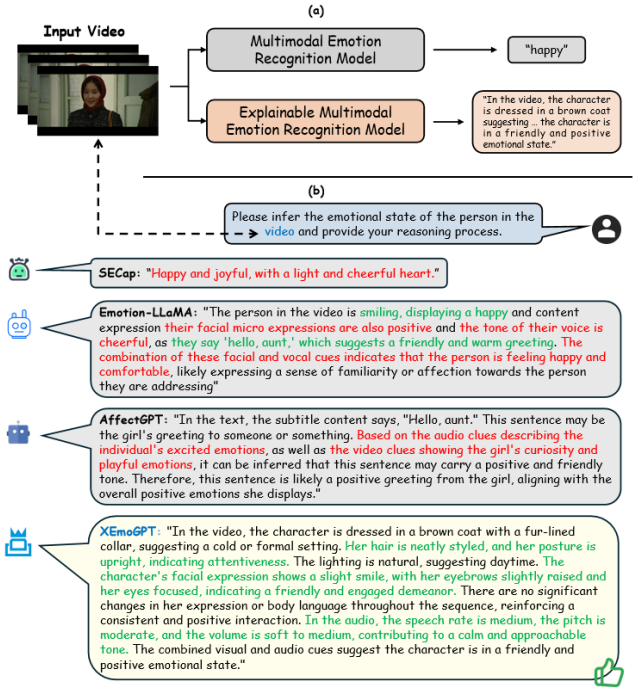

图 1. (a) 多模态情绪识别模型与可解释多模态情感识别模型的比较。(b) XEmoGPT 与其他情感多模态大语言模型的比较：绿色/红色文本表示带有/不带明确线索级解释的情绪预测。

I. 引言

情MOTION-AWARE技术在诸多现实应用中发挥着关键作用，包括人机交互 [1], 社交媒体分析 [2],和社交机器人学 [3]。随着这些系统日益参与直接影响人类用户的决策过程，仅仅预测情绪标签已不足够。相反，人们越来越需要模型能够通过透明且可解释的证据明确说明其预测依据，使用户能够理解为何推断出某种情绪状态。这一需求推动了从传统多模态情绪识别 (MER) [4],（侧重于结果级情绪分类）转向可解释多模态情感识别 (EMER) [5],（强调以证据为基础的情感推断），这是一种范式转变。如图 1(a) 所示，EMER 不仅预测情绪标签，还旨在揭示支持情感推断的底层感知与推理过程。

随着多模态大语言模型（MLLMs）的快速发展和广泛应用，近期的 EMER 方法利用其生成能力来生成情绪描述 [6]–[10], ，例如 SECap [11],、Emotion-LLaMA [12], 和 AffectGPT [13] 等模型。Emotion-LLaMA 基于经过指令微调的 LLaMA-2 [14], 构建，首先通过模态特定专家对音频、面部及上下文视觉线索进行编码，将其投影到共享的 token 空间，然后以自回归方式生成情绪标签及自由形式的文本解释；AffectGPT

Hanwen Zhang、Yao Liu、Peiyuan Jiang、Yihui He、Yajiao Deng、Siyu Du 和 Qiao Liu 任职于中国电子科技大学计算机科学与工程学院，地址为中国四川成都。Lang Junjie 和 Xie Jun 任职于中国电子科技集团有限公司第54研究所，地址为中国河北石家庄。本研究获得国家自然科学基金委员会项目 U22B2061 和四川省自然科学基金项目 2024NSFSC0496 的资助。通讯作者：Yao Liu 和 Qiao Liu。

则通过在单模态编码器与 Qwen-2.5 大语言模型 [15]，之间加入轻量级预融合模块（如 Q-Former 或基于注意力的池化），扩展了这一流程，从而在生成详细的情绪描述之前有效提炼跨模态交互。然而，尽管现有方法表现出色，其生成的描述大多以结果为导向，更像是对情绪的转述，而非真正的解释。如图 1(b) 所示，这类描述通常只说明预测出的情绪，至多笼统提及一些“线索”，却没有明确指出支撑推断的具体多模态证据。因此，这些解释缺乏透明度，也难以进行细粒度验证。

为解决这一问题，我们提出“情感线索”这一概念，将其定义为明确的、特定于模态的证据单元，用于衔接低层感知与高层情感推理。情感线索对应于一种可观察事实，它以一个或多个模态为依据，例如面部表情、身体姿态、声音语调或上下文物体，并可据此逻辑推断出情绪状态。例如，扬起的眉毛和温和的微笑构成了暗示快乐的视觉线索；而低垂的头部姿态与紧绷的面部表情相结合，则可能表示悲伤。通过将情感推理分解为由感知观察、线索识别、推理和情感推断组成的结构化层级，情感线索成为可解释的中间变量，明确连接低层感官信号与高层情感结论。

尽管在通用多模态理解方面取得了成功，现有基于多模态大语言模型的情感推理方法仍面临两项阻碍线索级推理的根本挑战：（1）通用模态编码器的预训练目标主要用于捕捉全局结构和通用语义表征，而非细粒度的情感线索，因此对情感信号的敏感度有限，如表 IV 所示。（2）如第 IV 节所述，现有数据集往往难以兼顾标注质量与规模：大型数据集缺乏精确的情感监督，而小型高质量数据集的数据规模有限，因而无法为情感推理提供充分的线索级指导。

为增强情感线索感知，我们提出 XE-moGPT，该模型集成了两个模块：视频情感线索桥（VECB）和音频情感线索桥（AECB）。VECB 通过位置编码和自注意力增强视觉编码器感知时序视觉情感线索的能力，并通过对比视频情感线索对齐、帧时序判别和掩码帧建模等辅助任务进一步训练。AECB 则通过对比音频情感线索对齐任务的训练，增强音频编码器对听觉情感线索的敏感度。

为实现细粒度情感推理，我们引入 EmoCue 数据集，为情感推理提供线索级监督。该数据集包括：（1）EmoCue-Instruct，通过模型主导、人工辅助 的策略 [13]，对 MER-Caption+ 进行细化，以注入明确的特定模态线索；以及（2）EmoCue-ShortCaption，它利用多模态大语言模型生成简洁的情感线索标注。

此外，当前 EMER 的评估指标很大程度上依赖基于提示的评分 [5],，这无法准确衡量模型生成文本与人工标注文本在情感线索层面的重合度。为解决这一局限，我们提出 EmoCue-360，这是一种新颖的自动评估指标，可从生成的描述中提取视觉、听觉和全局情感线索，并基于语义相似度计算线索级 F1 分数。为支持系统化评估，我们还发布了 EmoCue-Eval，这是一个由人工标注的基准，包含 400 个涵盖多种场景和情感状态的高质量样本。据我们所知，这是目前规模最大的专用于 EMER 的基准。我们的主要贡献总结如下：

- 我们提出了 XEmoGPT，并设计了 VECB 和 AECB 模块，通过模态特定的线索对齐来增强情感线索感知。大量实验表明，XEmoGPT 在传统生成指标和 EmoCue-360 上均取得了最先进性能，始终展现出在情感线索感知和推理方面稳健而全面的能力。

- 我们构建了 EmoCue 数据集，包含两部分：EmoCue-Instruct，含有 30k+个样本，并对情感状态和模态特定线索（视觉、音频和全局）进行了明确标注；以及 EmoCue-ShortCaption，包含 40k+ 条模型生成的简短情感线索。
- 我们设计了 EmoCue-360，这是一种自动评估指标，用于计算线索级 F1 分数。我们还发布了 EmoCue-Eval，这是规模最大的人工标注 EMER 基准，包含 400 个样本，可用于系统评估。

II. 相关工作

A. 多模态大语言模型

多模态大语言模型（MLLMs）在视觉问答 [16],、图像到文本生成 [17], 和跨模态推理 [18]等任务中获得了广泛关注。典型的多模态大语言模型架构由三个核心模块组成：预训练模态编码器、预训练大型语言模型（LLM）以及连接二者的模态接口 [6]。类比人类认知，模态编码器（例如用于图像的 CLIP-ViT [19] 或 EVA-CLIP [20] ，以及用于音频的 CLAP [21]）充当感觉系统，负责感知原始信号并将其预处理为紧凑表示。LLM 通常采用 LLaMA [14] 或 Vicuna [22]，等模型，是负责推理和生成响应的中央大脑。模态接口则将异构模态对齐到大语言模型能够理解的统一表示空间中。它可以通过可学习连接器实现，包括 BLIP-2 [23] 和 InstructBLIP [24]，中使用的 Q-Former，或 LLaVA [25],采用的 MLP 投影器；也可以通过将非文本输入转换为语言的专家模型实现。然而，现有多模态大语言模型在情感场景 [26], [27]中的能力仍然有限。即使是 GPT-4V [28]等先进模型，也难以识别音频情绪线索或细微的面部微表情，这主要是由于通用预训练目标与 EMER任务要求之间存在差距。

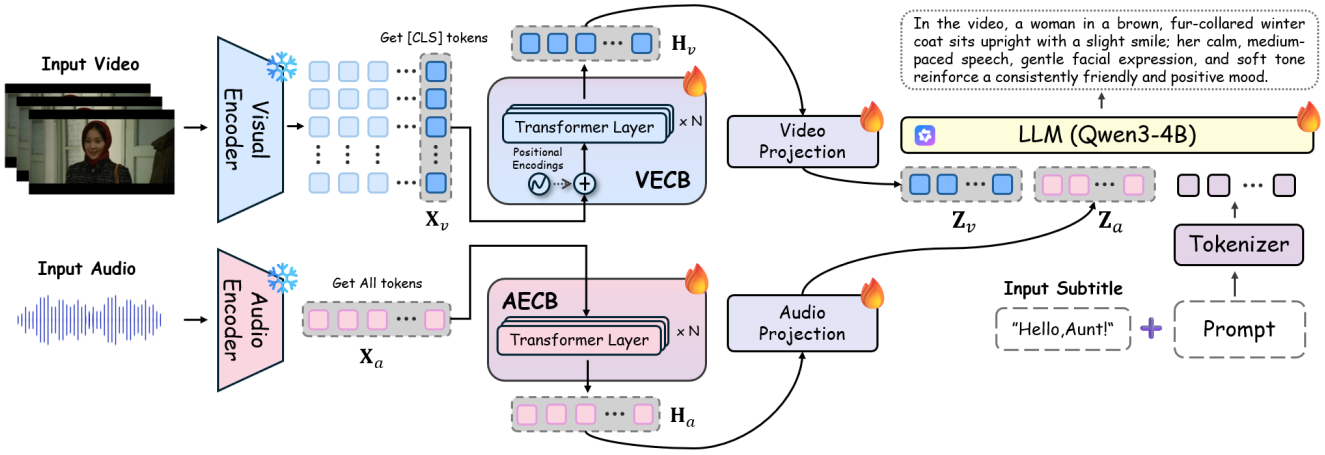

图2. XEmoGPT架构：该模型整合视觉、听觉和文本信息，生成包含视觉与听觉情感线索的描述。VECB和AECB模块主要用于增强模态编码器的情感线索感知能力。

B. 可解释多模态情感识别

可解释多模态情绪识别（EMER）任务旨在从视频、音频和文本模态中识别情感状态，同时为预测出的情绪提供明确且可解释的理由 [5]。近年来，情感计算领域的进展探索了针对情绪的多模态大语言模型微调 [7], [8], , 包括 SECap [11], 、Emotion-LLaMA [12], 和 AffectGPT [13],等模型，为 EMER 提供了富有前景的研究方向。然而，这些方法通常会生成以情绪为中心的描述，强调从视觉或听觉线索中推断出的情感状态，却没有明确指出支撑这些推断的具体多模态证据。

此外，现有用于EMER的自动评估指标在可靠性和可解释性方面仍存在局限。当前主流方法采用“基于提示的评分”[5]：将模型生成文本和人工标注文本连同精心设计的提示一起输入大语言模型，以生成评分。尽管这种指标简单易用，但仍存在一些局限：其评分过程可能受到提示设计以及大语言模型生成结果固有主观性的影响，从而影响对情感线索层面的对齐程度和信息重叠进行评估时的客观性与一致性。在一定程度上，这些因素可能会影响评估结果的可复现性和可靠性。

III. 方法

本节介绍 XEmoGPT 的架构，随后详细说明视频情感线索桥和音频情感线索桥的设计，最后介绍训练过程。

A. 模型架构

如图 2 所示，XEmoGPT 由四个与传统 MLLM 架构一致的组件构成，分别是视觉编码器、音频编码器、投影层和大语言模型。为增强情感线索感知能力，我们引入了两个新模块：视频情感线索桥（VECB）和音频情感线索桥（AECB）。

首先，我们使用冻结的 CLIP-ViT [19] 作为视觉编码器，并使用冻结的 HuBERT [29] 作为音频编码器，将视频帧和音频转换为一组视觉嵌入 $\mathbf{X}_v \in \mathbb{R}^{l_v \times d_v}$ 和音频嵌入 $\mathbf{X}_a \in \mathbb{R}^{l_a \times d_a}$，其中$l_v$ 表示视频帧的长度， l_a 表示音频嵌入的长度， d_v 和 d_a 分别表示视觉嵌入和音频嵌入的维度。接下来，我们使用 VECB 将视觉嵌入 $\mathbf{X}_v$ 转换为视觉情感嵌入 $\mathbf{H}_v \in \mathbb{R}^{l_v \times d_v}$，并使用 AECB 将音频嵌入 $\mathbf{X}_a$ 转换为音频情感嵌入 $\mathbf{H}_a \in \mathbb{R}^{l_a \times d_a}$。此外，我们使用视觉投影层和音频投影层，分别将 $\mathbf{H}_v$ 和 $\mathbf{H}_a$ 转换为视觉情感词元 $\mathbf{Z}_v \in \mathbb{R}^{l_v \times d}$ 和音频情感词元 $\mathbf{Z}_a \in \mathbb{R}^{l_a \times d}$。最后，将视觉情感词元 $\mathbf{Z}_v$、音频情感词元 $\mathbf{Z}_a$、视频字幕 T 和提示 P 输入大语言模型，以生成最终响应 $\hat{}\mathrm{R} = \mathrm{LLM}(\mathbf{Z}_v, \mathbf{Z}_a, \mathrm{T}, \mathrm{P})$。

1）视频情感线索桥：为区分不同时间步帧的顺序，VECB会向视觉嵌入中添加一组可学习的位置编码。随后，为了在时间维度上对视觉嵌入$\mathbf{X}_v$ 进行交叉融合，我们在VECB模块中引入Transformer，最终得到视觉情感嵌入$\mathbf{H}_v$。形式化地，这一过程可表示为

$$\mathbf{H}_v = \mathrm{Transformer}(\mathbf{X}_v + \mathrm{PE}(\mathbf{X}_v)) \tag{1}$$

其中，PE(·)表示位置编码。

2）音频情感线索桥：由于 HuBERT 能够输出细粒度、与时间步对齐的音频表征，并且已在内部引入位置编码来建模音频的时间信息，因此 AECB 模块无需添加额外的位置编码。相反，它直接使用 Transformer 在时间维度上对音频嵌入 $\mathbf{X}_a$ 进行交叉融合。该过程可表示为

$$\mathbf{H}_a = \mathrm{Transformer}(\mathbf{X}_a) \tag{2}$$

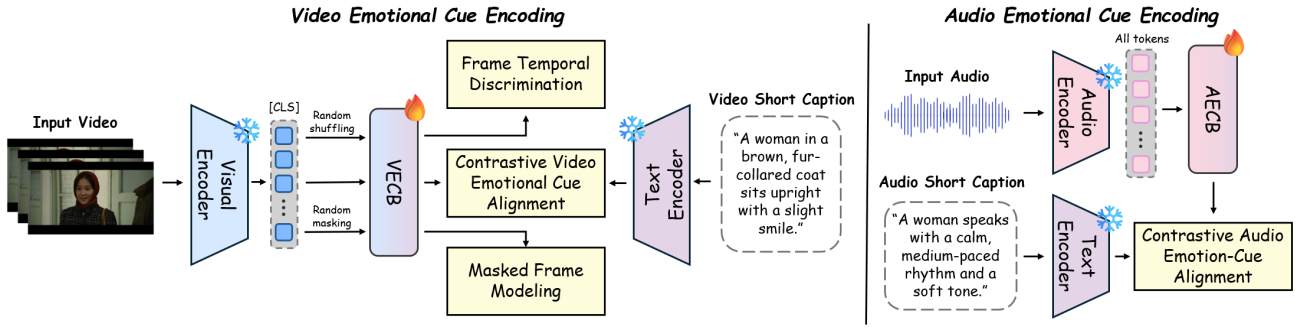

图3. VECB和AECB的训练过程：VECB模块通过三项辅助任务进行训练：对比视频情感线索对齐、帧时序判别和掩码帧建模。AECB模块通过对比音频情感线索对齐任务进行训练。

B. 视频情感线索编码

为增强预训练视觉编码器在线索级别的感知能力，并捕捉视频帧序列中的时间信息，我们将 VECB 模块与多项辅助任务联合训练，包括对比式视频情感线索对齐任务、帧时序判别任务和掩码帧建模任务，如图 3 所示。

1）对比视频情感线索对齐：为了弥合视觉嵌入与情感线索之间的差距，我们在VECB的输出与情感线索表示之间引入了对比学习目标。具体而言，我们首先沿时间维度对VECB输出$\mathbf{H}_v$ 进行池化，以获得全局视频表示$\mathbf{h}_v$。与此同时，使用冻结的CLIP文本编码器对情感线索进行编码，得到$\mathbf{h}_t$。为对齐视频与情感线索表示，我们采用InfoNCE损失 [30]，这是一种广泛使用的对比学习目标。它促使配对的视频和情感线索嵌入在表示空间中彼此接近，同时使非配对样本彼此远离：

$$\mathcal{L}_{\text{cl}}^{\text{vt}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_v, \mathbf{h}_t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{h}_v, \mathbf{h}_t^{(j)})/\tau)} \tag{3}$$

其中，$\text{sim}(\cdot,\cdot)$ 表示余弦相似度，τ 是可学习的温度参数，N 是批大小。

2）帧时序判别：为使 VECB 理解时序动态和帧的顺序，我们提出一项帧时序判别任务。给定一对视频片段，我们对帧序列进行采样并随机打乱其时序。随后要求模型预测该序列是否符合正确的时序顺序。我们在通过对 VECB 输出进行池化得到的全局视频表示$\mathbf{h}_v$ 之上应用一个简单的二元分类器：

$$\hat{y} = \text{MLP}(\mathbf{h}_v), \quad \mathcal{L}_{\text{temp}} = \text{BCE}(\hat{y}, y) \tag{4}$$

其中，$y \in \{0, 1\}$ 表示帧序列是否按顺序排列。

3）掩码帧建模：为了增强对帧的上下文理解，我们采用一种掩码帧建模策略，类似于 BERT 中的掩码语言建模 [31]。$\mathbf{X}_v$ 中的部分视觉嵌入会被随机掩码，并替换为一个可学习标记。随后，使用 VECB 进行训练以通过帧重建损失重建掩码位置处的原始帧嵌入：

$$\mathcal{L}_{\text{mfm}} = \sum_{i \in \mathcal{M}} \left\| \hat{\mathbf{x}}_v^{(i)} - \mathbf{x}_v^{(i)} \right\|_2^2 \tag{5}$$

其中，$\mathcal{M}$ 表示被掩码位置的集合，$\hat{\mathbf{x}}_v^{(i)}$是位置 i处预测得到的嵌入。

C. 音频情感线索编码

由于 HuBERT 是在大规模无监督音频数据上预训练的，其表示在音频空间中具备强大的上下文建模能力。然而，由于缺乏大规模音频-文本配对训练，其特征空间与情感线索的语义空间之间存在显著的语义差距。因此，如图 3 所示，AECB 的训练目标是通过将音频编码器的输出特征与情感线索的语义空间对齐，增强其情感线索感知能力。

1) 对比式音频情感线索对齐：首先，我们对音频情感嵌入 $\mathbf{H}_a$ 在时间维度上进行池化，得到全局音频表示 $\mathbf{h}_a$。情感描述由冻结的 LaBSE [32]编码，得到 $\mathbf{h}_t$。LaBSE 擅长多语言、多领域的句子相似度计算，因此适合对齐句子级情感线索。我们仍采用 InfoNCE 损失来对齐音频和情感线索：

$$\mathcal{L}_{\text{cl}}^{\text{at}} = -\log \frac{\exp(\text{sim}(\mathbf{h}_a, \mathbf{h}_t)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{h}_a, \mathbf{h}_t^{(j)})/\tau)} \tag{6}$$

其中，符号的含义与公式 3 中相同。

D. 训练过程

训练过程分为三个阶段。第一阶段着重训练 VECB模块，以获得视频情感线索表示。第二阶段训练 AECB模块，以获得音频情感线索表示。最后阶段将这些多模态表示与大语言模型的文本表示空间对齐。训练过程中使用的数据集列于表 I。

表I XEmoGPT三个训练阶段所用的训练数据集。阶段1和阶段2使用视频/音频-文本数据集；阶段3使用多模态指令微调数据集，以对齐情感推理。

阶段	数据集	Size
1	MAFW [34], DFEW [35], MER2025 赛道2 [36], EmoCue-ShortCaption	0.9M
2	SpeechCraft-Zhvoice [37]	10.2M
3	MAFW [34], EmoCue, MERCaption+ [13], SpeechCraft-50k [37]	1.6M

1）阶段*1*：训练 *VECB*：在此阶段，视觉编码器和文本编码器均被冻结，而 VECB模块则随机初始化。训练目标定义如下：

$$\mathcal{L}_{\text{VECB}} = \mathcal{L}_{\text{cl}}^{\text{vt}} + w_1 \cdot \mathcal{L}_{\text{temp}} + w_2 \cdot \mathcal{L}_{\text{mfm}} \tag{7}$$

其中，w_1 和 w_2 分别是帧时序判别损失和掩码帧建模损失的权重系数。

2) 阶段*2*：训练*AECB*：在此阶段，音频编码器和文本编码器均被冻结，AECB模块则随机初始化。训练目标定义如下：

$$\mathcal{L}_{\text{AECB}} = \mathcal{L}_{\text{cl}}^{\text{at}} \tag{8}$$

3) 阶段*3*：情感指令微调：在此阶段，CLIP和HuBERT编码器的参数保持冻结。对于大语言模型，我们应用LoRA [33] 以实现高效微调，使其能够以最低的计算开销适应EMER任务。相比之下，VECB和AECB模块以及音频和视觉投影层均可充分训练。

为增强模型基于模态特定线索进行推理的能力，我们在训练期间设计了三种类型的指令-响应对：(1) 仅视觉，(2) 仅音频，以及(3) 多模态（视觉和音频）。在优化过程中，我们采用定向梯度控制策略：对于仅视觉样本，阻断音频分支（包括AECB及其投影层）中的梯度；相反，对于仅音频样本，则阻断视觉分支中的梯度。这种选择性梯度流强化了模态特定归纳偏置，使模型能够在推理时根据用户指令关注相关模态，同时避免跨模态混淆。

IV. 训练数据集

目前，EMER任务的数据集面临样本量与标注质量之间的权衡。现有资源各有优势，且互为补充，但也存在局限。MERR- Coarse-Grained [12] 提供了一个大规模数据集（样本数 N = 28,618），其标注粒度较粗；而 MERR-Fine-Grained [12]则提供了更详细的标注，但规模较小（N = 4,487）。MER-Caption+ [13]在样本量（N =31,327）与标注质量之间实现了更均衡的折中，但其主要关注文本情感描述，未明确标注视觉或听觉线索。

为克服这些局限，我们提出了 EmoCue，这是一个由两部分组成的新数据集：EmoCue-Instruct（N = 31,327）和 EmoCue-ShortCaption（N = 40,683）。

EmoCue-Instruct 基于 MER-Caption+，采用模型主导、人工辅助 的策略重写其标注：(1) 向多模态大语言模型提供 MER2025 [36]情绪标签作为提示，以推断模态特定情感线索；(2) deepseek-chat [38], [39] 整合这些线索，确保一致性并解决冲突；(3) 专家随机选取 300 个样本并进行人工评分，以评估数据集质量，如第 VI-F 节所示。

EmoCue-ShortCaption 基于 DFEW [35]和 MER-Caption+构建，利用多模态大语言模型生成简洁的情感线索，无需人工参与。

V. 评估指标与数据集

如第二节所述，当前 EMER 任务的评估方法仍有局限。为解决这些问题，受 [40]，启发，我们提出了一种全新的自动评估指标 EmoCue-360，可在情感线索层面精确量化模型生成文本与人工标注文本之间的重合度。

A. 任务定义

我们首先对 EMER 任务作出更详细的定义：给定一个视频片段，模型需要生成一段可解释的情感描述 $\mathcal{D}$，其语义层面应包含以下三部分信息：

$$\mathcal{D} = \{\mathcal{C}_v, \mathcal{C}_a, \mathcal{C}_s\} \tag{9}$$

其中，$\mathcal{C}_v$、$\mathcal{C}_a$ 和 $\mathcal{C}_s$ 分别表示视觉、音频和全局情感线索集合。每个线索集合由若干原子情感线索 c 组成。这些原子线索可视为文本中最小的、语义独立且可辨识的单位。理论上，任何自然语言文本都可以分解为此类原子线索。 [41]

给定一对模型生成文本 $\mathcal{D}_{\text{ca}}$ 和人工标注文本 $\mathcal{D}_{\text{gt}}$，高质量 $\mathcal{D}_{\text{ca}}$的每个维度都应满足以下两个条件：

- **完整性**：出现在 $\mathcal{D}_{gt}$ 某一维度中的任何原子线索，都应在 $\mathcal{D}_{ca}$的相应维度中出现语义相同的原子线索。
- **正确性**：出现在 $\mathcal{D}_{ca}$ 某一维度中的任何原子线索，都应在 $\mathcal{D}_{gt}$的相应维度中出现语义相同的原子线索。

B. 指标计算

EmoCue-360 指标的计算流程如图 4 所示。给定一对模型生成文本$\mathcal{D}_{\text{ca}}$ 和人工标注文本 $\mathcal{D}_{\text{gt}}$，我们首先需要从文本中提取上述三个维度的情感线索集合。为解决开放式文本的线索提取问题，我们可以利用大语言模型在

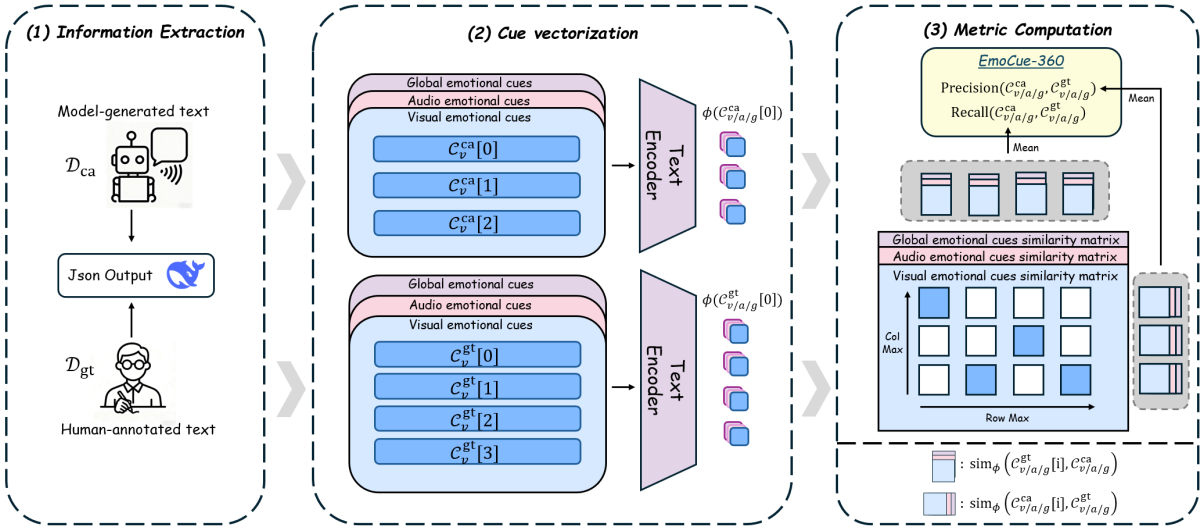

图4. EmoCue-360 指标的计算流程：信息抽取、线索向量化和指标计算。所用数学公式与第V节中的公式相对应。

信息抽取任务中的强大能力 [42]，自动将开放式文本分解为经过细化的线索列表。

接下来，对于 $\mathcal{D}_{\text{ca}}$ 和 $\mathcal{D}_{\text{gt}}$的每个维度，我们都可以计算精确率和召回率。以视觉情感线索为例，模型生成文本中的视觉情感线索集合为 $\mathcal{C}_v^{\text{ca}}$，人工标注文本中的视觉情感线索集合为 $\mathcal{C}_v^{\text{gt}}$。因此，精确率和召回率定义如下：

$$\text{Precision}(\mathcal{C}_v^{\text{ca}}, \mathcal{C}_v^{\text{gt}}) = \frac{1}{|\mathcal{C}_v^{\text{ca}}|} \sum_{i=1}^{|\mathcal{C}_v^{\text{ca}}|} \text{sim}_\phi(\mathcal{C}_v^{\text{ca}}[i], \mathcal{C}_v^{\text{gt}}) \tag{10}$$

$$\text{Recall}(\mathcal{C}_v^{\text{ca}}, \mathcal{C}_v^{\text{gt}}) = \frac{1}{|\mathcal{C}_v^{\text{gt}}|} \sum_{i=1}^{|\mathcal{C}_v^{\text{gt}}|} \text{sim}_\phi(\mathcal{C}_v^{\text{gt}}[i], \mathcal{C}_v^{\text{ca}}) \tag{11}$$

其中，$\text{sim}_\phi(\mathcal{C}[i],\ \mathcal{C}') = \max_{\mathcal{C}'[j]\in\mathcal{C}'} \cos(\phi(\mathcal{C}[i])\ \phi(\mathcal{C}'[j]))$ 表示线索 $\mathcal{C}[i]$ 与线索集合 $\mathcal{C}'$之间的相似度；$\phi(\cdot)$是一个句子级文本编码器，可将文本线索映射为向量表示；$\cos(\cdot,\ \cdot)$ 表示两个向量之间的余弦相似度。为平衡精确率和召回率，我们将它们的调和平均作为 F_1 分数。

C. 评估数据集

EmoCue-360的一个前提假设是，用于评估的人工标注数据必须具有足够高的质量，并能够覆盖视频片段中存在的所有主要情绪线索。目前，在公开可用的数据集中，只有EMER数据集 [5]能够在标注质量和线索级粒度方面合理地满足这一要求。为进一步促进EMER任务中的系统性评估研究，我们额外构建了一个新的基准EmoCue-Eval，其中包含400个经过仔细精细标注的测试样本，涵盖多样的场景、说话人和情感状态。据我们所知，EmoCue-Eval是目前同类中规模最大的EMER基准。该数据集为未来EMER模型比较研究提供了更为系统的基准。

VI. 实验

我们开展了一系列实验，以评估XEmoGPT的有效性，并解决以下研究问题：

- **RQ1**：与现有方法相比，XEmoGPT在EMER任务上的表现如何?
- **RQ2**：EmoCue-360对提示模板和文本风格变化的鲁棒性如何?
- **RQ3**：VECB和AECB能否增强通用模态编码器的情感线索感知能力?
- **RQ4**：VECB和AECB能否提升XEmoGPT的情感线索推理能力?
- **RQ5**：EmoCue-Instruct 数据集的质量是否足以支持模型训练?

A. 实现细节

我们的实验在一台配备两张 NVIDIA A100 GPU（每张显存 80GB）的机器上进行，并采用 BF16格式的混合精度训练以提升计算效率。对于多模态输入处理，每个视频均匀下采样为 16 个具有代表性的关键帧，空间分辨率为 224×224，随后进行 CLIP式归一化。音频波形以 16 kHz 的采样率处理，并从中提取 HuBERT特征。三个训练阶段的超参数设置如下：（1）VECB模块训练 30 个训练轮次，批大小为 512，学习率为 3×10^{-4}；整个训练过程中，CLIP-ViT [19] 视觉特征保持冻结；（2）AECB模块采用相近的设置进行训练，批大小为 448，并冻结 HuBERT-Chinese [29] 音频特征；（3）在指令微调阶段，我们采用 Qwen3-4B [51]作为骨干大语言模型，并使用 LoRA [33] 进行适配，秩设为 512。此阶段将批大小降至 8，学习率设为 1×10^{-5}。

表 II EMER 和 EmoCue-Eval 基准数据集上的性能比较（A：音频，V：视频）。评估指标包括 BLEU、CIDEr，以及视觉（VIS-EMO）、音频（AUD-EMO）和全局（GLO-EMO）情感线索的 EmoCue-360 分数，其中精确率（P）和召回率（R）的平均值为所有 P/R 分数的平均值。**BOLD** 表示最佳结果，下划线表示次佳结果。

模型	模态	大小 ↓	$BLEU_1$ ↑	$BLEU_4$ ↑	CIDEr ↑	VIS-EMO		AUD-EMO		GLO-EMO		平均值 ↑
						P ↑	R ↑	P ↑	R ↑	P ↑	R ↑	
EMER												
通用多模态大语言模型												
Qwen2-Audio [43]	A	7B	0.0848	0.0086	0.0004	-	-	38.4	41.8	47.9	42.5	42.6
SALMONN [44]	A	7B	0.2297	0.0170	0.0018	-	-	43.5	48.4	45.2	41.2	44.6
LLaVA-NeXT [25]	V	7B	0.2208	0.0144	0.0072	48.8	46.2	-	-	46.7	44.8	46.6
Qwen2.5-VL [45]	V	3B	0.1950	0.0234	0.0075	42.0	33.7	-	-	48.7	46.4	42.7
LLaMA-VID [46]	V	7B	0.2165	0.0311	0.0034	47.7	49.1	-	-	45.4	42.2	46.1
mPLUG-Owl [47]	V	7B	0.2175	0.0207	0.0037	43.8	47.9	-	-	45.3	43.0	45.0
Video-ChatGPT [48]	V	7B	0.2478	0.0369	0.0023	45.1	46.7	-	-	45.0	41.9	44.7
VideoChat [49]	V	7B	0.2542	0.0254	0.0055	42.6	48.6	-	-	42.4	42.7	44.1
VideoChat2 [50]	V	7B	0.1251	0.0155	0.0001	41.4	36.8	-	-	46.7	39.2	41.0
情感多模态大语言模型												
SECap [11]	A	7B	0.0001	0.0000	0.0000	-	-	28.7	29.4	43.0	35.4	34.1
Emotion-LLaMA [12]	V+A	7B	0.1758	0.0289	0.0017	**51.1**	39.1	39.2	46.7	49.5	46.9	45.4
AffectGPT [13]	V+A	7B	0.2418	**0.0691**	0.0029	34.2	27.5	**46.1**	46.8	**50.9**	48.0	42.3
XEmoGPT（本文方法）	V+A	4B	**0.2965**	0.0655	**0.0349**	44.7	**51.7**	38.2	**50.3**	50.4	**49.6**	**47.5**
EmoCue-Eval												
情感多模态大语言模型												
SECap [11]	A	7B	0.0001	0.0000	0.0000	-	-	26.5	22.1	39.1	30.6	29.6
Emotion-LLaMA [12]	V+A	7B	0.1326	0.0102	0.0000	**49.9**	39.5	37.4	36.7	41.7	37.8	40.5
AffectGPT [13]	V+A	7B	0.1368	0.0076	0.0000	32.7	25.1	42.1	34.9	39.2	35.9	35.1
XEmoGPT（我们的方法）	V+A	4B	**0.2430**	**0.0488**	**0.0003**	43.2	**48.7**	**49.2**	**49.3**	**45.1**	**41.6**	**46.2**

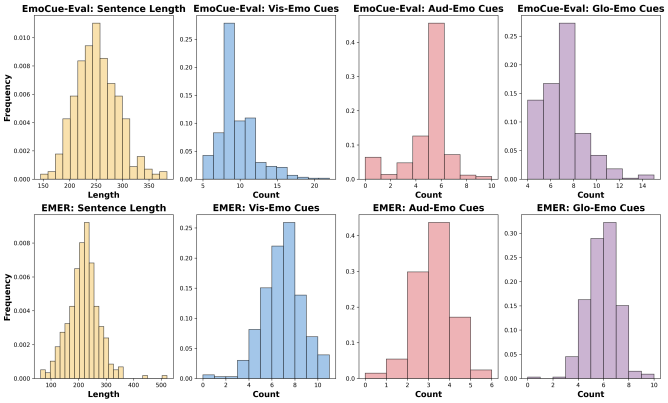

图 5. EmoCue-Eval 与 EMER 数据集的质量对比，展示了标注长度（第 1 列）以及情感线索数量的分布（第 2–4 列）。

B. 零样本评估（*RQ1*）

为展示 XEmoGPT 在感知情感线索并进行推理方面的强大能力，我们在两个基准数据集上开展零样本评估：EMER数据集和 EmoCue-Eval。评估指标既包括 BLEU [52] 和CIDEr [53], 等传统文本生成指标，也包括我们提出的 EmoCue-360，用于评估情感线索层面的表现。

如表 II 所示，XEmoGPT 在 EMER数据集上取得了最高的平均分，优于所有基线模型。它在 BLEU-1 和 CIDEr 上表现领先，并且在 EmoCue-360 的所有情感线索类型中均取得了最高召回率。尽管其在 EMER 上的精确率分数相对较低，这种表现可以用 EMER 标注的特性来解释，而非模型本身的缺陷。

具体而言，图 5 从标注长度和情绪线索密度两方面详细比较了 EMER 与 EmoCue-Eval。图中显示，EMER 的标注明显更短，每个样本包含的明确情感线索也更少。在许多情况下，EMER 中的情绪描述主要关注高层次的情绪状态，而省略了细粒度的视觉或听觉证据。因此，使用 EmoCue-360 等线索级指标进行评估时，XEmoGPT 往往会生成额外有效且具有感知依据的情感线索，而这些线索并未出现在 EMER 的标注中。这些额外线索因而被计为假阳性，导致精确率分数较低，尽管它们在语义上是合理的。

相比之下，EmoCue-Eval 在视觉、听觉和全局维度上提供了全面且密集的线索标注，大幅降低了标注稀疏性。在这一设置下，XEmoGPT 在几乎所有指标上都始终优于所有基线方法，包括精确率和召回率。尽管 Emotion-LLaMA 在视觉线索精确率上略胜于 XEmoGPT，但其召回率和全局线索推理表现较差，这表明它采用了更为保守但不够全面的解释策略。

C. EmoCue-360 鲁棒性分析（*RQ2*）

本实验评估 EmoCue-360 指标的鲁棒性。具体而言，我们从两个方面评估其鲁棒性：对提示模板的独立性，以及对文本风格的不敏感性。

1) 不依赖提示模板：我们设计了五种不同的提示模板，并使用 deepseek-chat 从

表 III VIS-EMO（视觉）、AUD-EMO（音频）和 GLO-EMO（全局）情感线索的精确率（P）、召回率（R）和 F1 分数的标准差（标准差）及范围。所有数值均按 10^{-3} 缩放。

	VIS-EMO			AUD-EMO			GLO-EMO		
	P	R	F1	P	R	F1	P	R	F1
标准差 ↓	3.1	3.4	2.3	1.5	2.3	1.6	0.9	2.3	2.1
范围 ↓	7.1	8.9	5.3	3.6	5.6	4.1	2.3	5.8	5.3

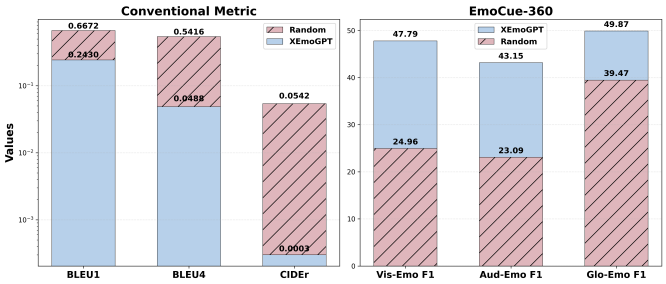

图6. 传统指标与 EmoCue-360 的比较："随机"表示人工构造的样本对，而"XEmoGPT"表示其在 EmoCue-Eval 数据集上的得分。

EMER数据集中沿三个维度提取原子级情感线索。基于每种模板生成的线索，我们计算了 XE-moGPT 的 EmoCue-360 分数。如图7所示，EmoCue-360 对提示设计的变化表现出稳定的鲁棒性。具体而言，表 III 显示，所有评估维度的范围和标准差均稳定在 10^{-3} 的量级。

2) 对文本风格的不敏感性：我们设计了一个模板，并构建了一组涵盖视觉和音频模态的情感线索。每次试验中，我们随机将部分线索插入模板，合成两个文本片段，分别模拟人工标注和模型输出，共得到 50 对样本。随后，我们计算了这些样本对的传统文本生成指标和 EmoCue-360 得分。如图6所示，对于合成样本，传统指标得分始终高于 XEmoGPT 在 EmoCue-Eval 上的表现。相比之下，使用 EmoCue-360 时，XEmoGPT 的得分始终高于合成样本。这表明，与传统指标相比，EmoCue-360 更注重情感线索的一致性，而非生成文本整体的风格相似性。

D. 与通用编码器的比较（*RQ3*）

为验证所提出的VECB和AECB模块在增强通用模态编码器情感线索感知能力方面的有效性，我们在两个具有代表性的情感识别基准数据集M3ED和MER2023Track1上进行了评估。这两个基准数据集在数据分布、标注方案和模态特征方面存在显著差异，从而为评估线索增强编码器的泛化能力提供了全面的测试平台。我们报告非加权平均召回率（UAR）

表 IV 基于 UAR 与 WAR 的 M3ED 和 MER2023 TRACK1 模态编码器性能比较。**粗体**表示最佳，下划线表示次佳。

模型	模态	M3ED		MER2023	
		UAR	WAR	UAR	WAR
CLAP [21]	A	18.3	4.5	17.5	14.1
LanguageBind [54]	A	17.5	16.6	19.6	10.9
CLIP [19]	V	22.6	23.7	30.3	29.7
LanguageBind [54]	V	16.4	11.7	28.1	28.0
LongCLIP [55]	V	23.0	23.8	39.1	41.1
VideoCLIP-XL [56]	V	20.7	24.7	32.6	35.0
HuBERT+AECB	A	**26.4**	32.2	40.4	43.6
CLIP+VECB	V	24.8	**33.0**	**42.6**	**43.8**

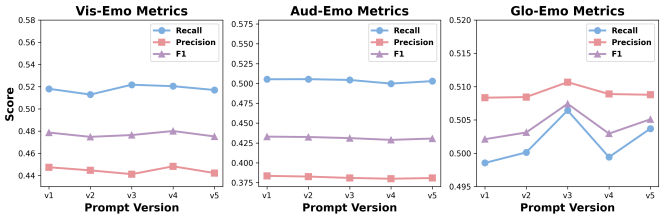

图 7. EmoCue-360 指标在 5 种提示模板下的稳定性，展示了视觉、音频和全局情感线索的一致性能。

和加权平均召回率（WAR），以兼顾类别不平衡和整体识别准确率。

如表 IV 所示，整合 VECB 和 AECB 始终能显著提升相较于相应骨干编码器的性能。与视觉模态下 CLIP、LongCLIP 和 VideoCLIP-XL 等强大的通用基线相比，CLIP+VECB 在两个数据集上均表现更优，这表明，显式建模情感线索比单纯扩大模型规模或延长时间上下文更能产生判别性表征。具体而言，CLIP+VECB 在 MER2023 上取得了最佳整体性能，UAR 达到 42.6，WAR 达到 43.8。

对于听觉模态，配备 AECB 的 HuBERT 显著优于 CLAP 和 LanguageBind 等仅音频基线。值得注意的是，HuBERT+AECB 在 M3ED 上取得最高 UAR，这表明，在情感信号较为微弱且分布于各种声学模式中的场景下，具备线索感知能力的音频表征尤为有效。其在 MER2023 上的竞争力进一步表明，AECB 能够很好地适应录音条件和标签分布各异的数据集。

这些结果凸显了通用多模态编码器与线索增强型编码器之间的重要区别。现有编码器擅长捕获全局语义信息，但并未专门针对情感显著模式进行优化。通过引入 VECB 和 AECB，编码器能够更专注于细粒度且与情感相关的区域和声学特征，从而获得更稳健、更具迁移性的情感表征。



Ground Truth: "In the video, we see a lady using a phone in an indoor environment. Her facial expression appears joyful, with the corners of her mouth turned up, indicating that she may be hearing some good news that pleases her. As time goes on, her eyes slightly squint, but her smile remains evident, accompanied by a slight nod, which is a natural reaction when people hear good news or something interesting. In the audio, the character's voice is loud, with a calm tone and a hint of laughter, indicating that the character is in a pleasant mood. In the text, the subtitle says, "Then let's meet, okay? Well, goodbye." This sentence may be an expression of the lady making an appointment to meet the other person over the phone. Based on the lady's joyful facial expression and smile in the video, as well as the character's pleasant tone and laughter in the audio, we can infer that this sentence is a positive response from the lady to the other person, indicating her willingness to meet. The lady's smile and nod further support this inference. Therefore, this sentence expresses the lady's happiness and anticipation."

SECap: "This person cares about other people's feelings and feels quite bored."

Emotion-LLaMA: "The person in the video displays a happy and relaxed facial expression, with their eyebrows slightly raised and a smile on their face their tone is also cheerful, as they say 'oh, okay, goodbye' in a friendly manner the combination of their facial expression and tone suggests that they are expressing a positive emotion, likely happiness or contentment."

AffectGPT: "In the text, the subtitle reads: "Be there or be square. Oh, okay, goodbye." Based on the description of the individual's calm and steady tone of voice in the audio clue, as well as the serious facial expression and body language of the man in the video clue, we can infer that this sentence may be a response from the man to the woman. Due to the neutral tone of the individual in the audio clue, as well as the serious expression and body language of the man in the video clue, we can speculate that this sentence may have a neutral or indifferent tone. Therefore, this sentence may not be an expression of emotion, but rather a neutral response or an answer to the woman's question."

XEmoGPT: "In the video, a woman is engaged in a phone conversation in a cozy, well-decorated room with soft lighting, creating a warm and intimate atmosphere. She wears a light-colored blazer over a patterned top, suggesting a professional yet casual setting. Her facial expressions are animated, with raised eyebrows, wide eyes, and a slightly open mouth, conveying enthusiasm and engagement. Her body language is relaxed yet attentive, leaning slightly forward as she speaks, reinforcing her active participation in the conversation. In the audio, the speaker's speech rate is fast, with fluctuating intonation and an assertive vocal tone. The pitch varies dynamically, and the volume is moderately loud, contributing to an energetic and confident delivery. In the text, the dialogue includes phrases like "Be there or be square" and "Okay, goodbye," which, combined with the speaker's assertive tone, suggest a sense of decisiveness and positivity. The integration of visual, auditory, and textual cues indicates a highly positive emotional state, characterized by enthusiasm, engagement, and confidence.."

图8. EMER案例研究：样本00000007的样本对比

表 V
模态特定桥接模块的消融结果。“VIS-EMO”“AUD-EMO”和“GLO-EMO”分别表示视觉、音频和全局情感线索的 F1 分数。VECB 和 AECB 指视觉和音频情感线索桥接模块，它们能够增强 XEmoGPT 的模态特定情感线索推理。

模块		VIS-EMO	AUD-EMO	Glo-Emo
VECB	AECB	F1 ↑	F1 ↑	F1 ↑
-	-	46.941	42.325	47.853
✓	-	<u>47.582</u>	42.465	<u>48.764</u>
-	✓	46.361	<u>43.125</u>	48.372
✓	✓	**47.785**	**43.148**	**49.866**

E. 消融研究（*RQ4*）

表 V 展示了消融研究，评估了模态特定桥接模块各自及协同发挥的作用。结果揭示了若干重要发现，说明 VECB 和 AECB 如何促进 XEmoGPT 中的情感线索感知与推理。

首先，每个桥接模块都表现出明确的模态特异性有效性。引入VECB可显著提升视觉情感线索识别，将视觉情感 F1 分数从46.941提高到47.582，同时音频线索表现基本保持不变。同样，引入AECB主要增强听觉情感线索推理，将听觉情感 F1 分数从42.325提高到43.125，对视觉线索仅有轻微影响。这些结果证实，这两个模块并非通用的性能提升器，而是有选择地强化各自模态内与情感相关的表征。

其次，同时启用 VECB 和 AECB 时，XEmoGPT 在包括全局情感线索在内的所有三个维度上均取得最佳表现。值得注意的是，Glo-Emo 的 F1分数从基线模型的 47.853 提升至同时应用两个模块时的 49.866。这一提升超过了单独引入任一模块所带来的增益，表明视觉与听觉情感线索之间存在互补效应。通过增强模态特定的线索感知，模型能够更好地开展整体情感推理，整合跨模态证据。

总体而言，这些消融结果验证了

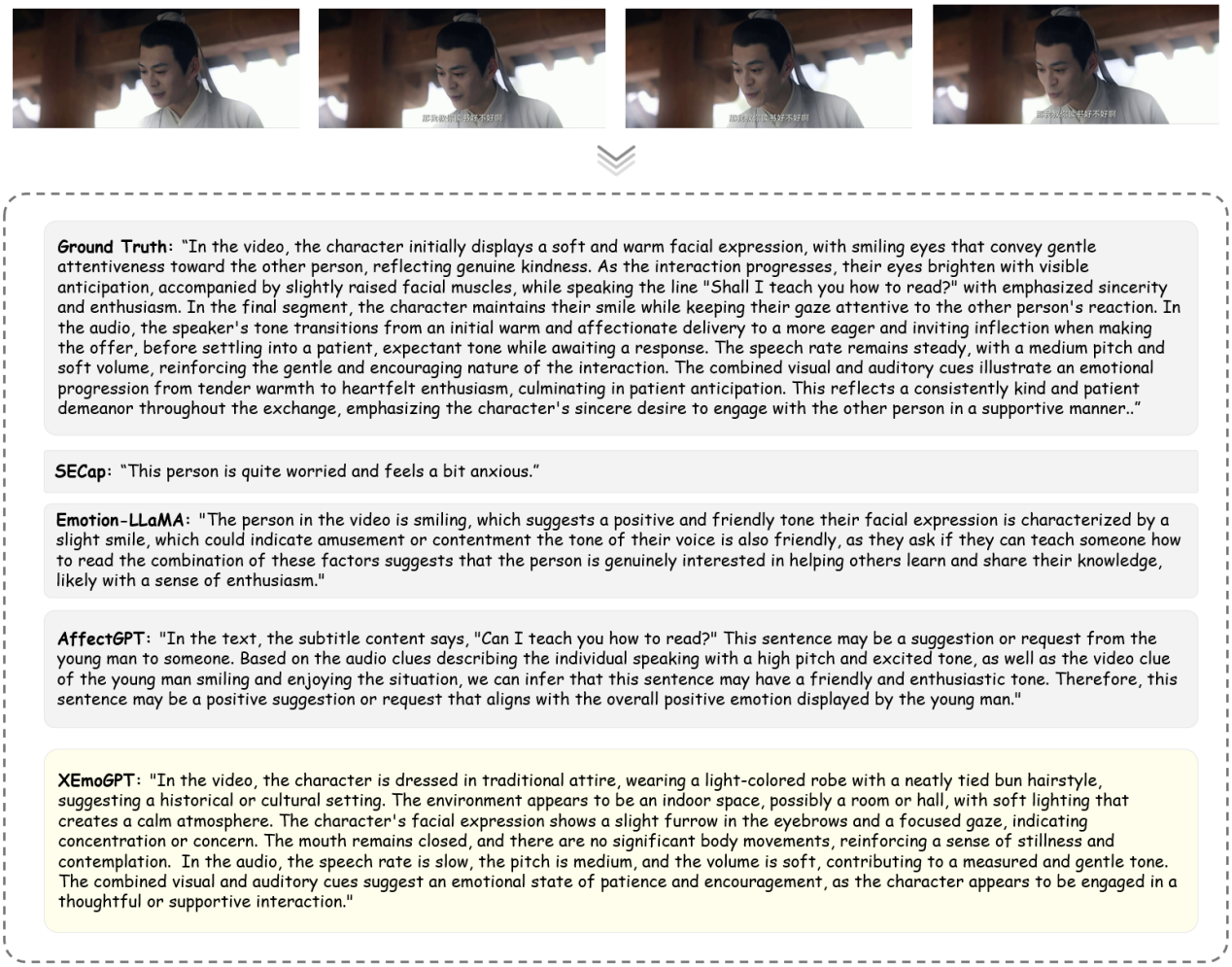

图 9. EmoCue-Eval 案例研究：samplenew3 00070374 的样本比较 _

VECB 和 AECB 的设计动机。通过在模态层面显式建模情感线索，所提出的桥接模块为情感推理提供了更具结构性且可解释的路径，最终带来更准确、更全面的情感解释。

这些定量发现也得到了图 8 所示案例研究中定性证据的进一步支持。在该案例中，VECB 和 AECB 的联合使用使 XEmoGPT 能够在多种模态中识别出更完整、连贯且以感知为依据的情感线索。

F. EmoCue-Instruct 质量评估（*RQ5*）

.

如第四节所述，EmoCue-Instruct 数据集中的情感描述主要由多模态大语言模型根据情感标签进行逆向推理生成，随后使用 DeepSeek-Chat 进行摘要。由于这一过程没有人工干预，因此有必要进行人工评估，以评定生成数据的标注质量。

相比之下，EmoCue-ShortCaption 数据集的质量已通过 VECB 和 AECB 在表 IV 中的优异表现得到间接验证。因此，并且

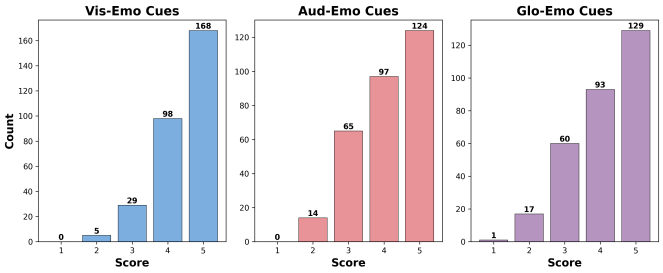

图 10. EmoCue-Instruct 数据集的人工评估得分分布。得分为三位专家在视觉、音频和全局情感线索维度上的平均值。每项得分范围为 1（差于 MER-Caption+）至 5（优于 MER-Caption+）。

考虑到人工评估成本较高，我们未对该数据集进行额外的人工评估。

1）评估方案：我们从 EmoCue-Instruct 数据集中随机抽取 300 个样本。由于该数据集改编自 MER-Caption+，我们在 MER-Caption+ 中找到相应的标注，作为人工标注参考。我们邀请了三位具有情感计算研究经验的专家

为每个样本进行评分，比较两个数据集的标注质量。评估聚焦于三个维度：视觉情感线索、音频情绪线索和全局情感线索。每个维度采用 1–5 分制进行评定：1 分表示 EmoCue-Instruct 的标注质量低于 MER-Caption+，3 分表示两个数据集的质量相当，5 分表示 EmoCue-Instruct 提供的标注优于 MER-Caption+。

如图 10 所示，EmoCue-Instruct 在全部三个维度上均取得了更高的平均分，表明其自动生成的标注保持了较高质量，适合用于训练可解释多模态情感识别模型。

G. 案例研究

本节基于两个数据集展示案例研究：EMER 和 EmoCue-Eval。图 8 展示了 EMER数据集中的一个代表性样本，而图 9 展示了 EmoCue-Eval 中的一个样本。每个案例都包含一个视频帧，以及由各种情绪多模态语言模型生成的相应字幕。

图 8 展示了来自 EMER 的一个代表性示例。与基线方法相比，XEmoGPT 提供的解释更明确地基于多模态情绪线索。SECap 给出了错误的高层情绪判断，而 Emotion-LLaMA 和 AffectGPT 主要提供粗粒度的情绪摘要；相比之下，XEmoGPT 识别出具体的视觉和听觉线索，并基于这些观察进行推理，以推断细微的情绪状态，从而与真实标注更为一致。

图 9 展示了 EmoCue-Eval 中一个带有密集线索标注的样例。在此案例中，XEmoGPT 能更有效地整合视觉、听觉和文本线索，推断出积极且投入的情绪状态；而基线方法要么给出过于笼统的描述，要么误解情绪基调。这些示例从定性角度展示了 XEmoGPT 在线索级情感推理方面的优势。

VII. 结论

我们提出了 XEmoGPT，这是一种新型多模态大语言模型，通过感知跨模态的情感线索，推动了可解释情感推理的发展。为加强线索级感知，我们提出了 VECB 和 AECB，以增强模态编码器并为多模态情感推理奠定基础。此外，我们构建了 EmoCue 数据集，为线索级情感推理提供细粒度监督。为实现可解释评估，我们开发了 EmoCue-360，用于线索级评估，并发布了 EmoCue-Eval，这是迄今规模最大的 EMER 人工标注基准。实验结果表明，XEmoGPT 在跨模态情感线索感知与推理方面表现出色。

未来的工作将着重探索线索—情绪推理的内在机制，因为当前模型依赖大量噪声数据来学习这类推理链。一个关键方向是研究如何有效过滤掉虚假线索，并实现更加可靠的线索—情绪推理。潜在方向包括采用专门针对情绪的编码器，改进 EmoCue-360 的句子级文本编码器，从而实现更敏锐且更具上下文感知能力的线索匹配；此外，还包括优化 EmoCue 数据集的标注策略，进一步提升训练监督的质量和一致性。

参考文献

[1] J. Preece、Y. Rogers、H. Sharp、D. Benyon、S. Holland 和 T. Carey，《人机交互》。Addison-Wesley Longman Ltd.，1994年。[2] B. Batrinca 和 P. C. Treleaven，“社交媒体分析：技术、工具与平台综述”，《*Ai & Society*》，第30卷，第1期，第89–116页，2015年。[3] C. Breazeal、K. Dautenhahn 和 T. Kanda，“社交机器人学”，《*Springer* 机器人学手册》，第1935–1972页，2016年。[4] N. Ahmed、Z. Al Aghbari 和 S. Girija，“基于学习算法的多模态情绪识别系统综述”，《*Intelligent SystemswithApplications*》，第17卷，文章编号200171，2023年。[5] Z. Lian、L. Sun、M. Xu、H. Sun、K. Xu、Z. Wen、S. Chen、B. Liu 和 J. Tao，“可解释的多模态情绪推理”，《*CoRR*》，2023年。[6] J. Wu、W. Gan、Z. Chen、S. Wan 和 P. S. Yu，“多模态大语言模型：综述”，载于*2023*年 *IEEE* 国际大数据会议（*BigData*）。IEEE，2023年，第2247–2256页。[7] H. Xie、C.-J. Peng、Y.-W. Tseng、H.-J. Chen、C.-F. Hsu、H.-H. Shuai 和 W.-H. Cheng，“Emovit：通过视觉指令微调革新情绪洞察”，载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，2024年，第26 596–26 605页。[8] Z. Liang、H. Shi 和 H. Chen，“AlignCap：使语音情绪描述与人类偏好保持一致”，载于*2024*年自然语言处理实证方法会议论文集，Y. Al-Onaizan、M. Bansal 和 Y.-N. Chen 编。美国佛罗里达州迈阿密：计算语言学协会，2024年11月，第3837–3846页。[在线]。获取地址：https://aclanthology.org/2024.emnlp-main.224/ [9] Z. Lian、F. Zhang、Y. Zhang、J. Tao、R. Liu、H. Chen 和 X. Li，“Affectgpt-r1：利用强化学习实现开放词汇多模态情绪识别”，2025年。 [在线]。获取地址：https://arxiv.org/abs/2508 01318 。F. Zhang、H. Li、S. Qian、X. Wang、Z. Lian、H. Wu、Z. Zhu、Y. Gao、Q. Li、Y. Zheng、Z. Lin 和 P.-A. Heng，“多模态大语言模型时代的面部表情识别再思考：基准、数据集及其他”，2025年。 [在线]。获取地址：https://arxiv.org/abs/2511 00389。Y. Xu、H. Chen、J. Yu、Q. Huang、Z. Wu、S.-X. Zhang、G. Li、Y. Luo 和 R. Gu，“SECap：使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，第38卷，2024年，第19 323–19 331页。[12] Z. Cheng、Z.-Q. Cheng、J.-Y. He、K. Wang、Y. Lin、Z. Lian、X. Peng 和 A. Hauptmann，“Emotion-LLaMA：通过指令微调进行多模态情绪识别与推理”，《神经信息处理系统进展》，第37卷，第110 805–110 853页，2024年。[13] Z. Lian、H. Chen、L. Chen、H. Sun、L. Sun、Y. Ren、Z. Cheng、B. Liu、R. Liu、X. Peng、J. Yi 和 J. Tao，“AffectGPT：用于借助多模态大语言模型理解情绪的新数据集、模型和基准”，载于第四十二届国际机器学习会议，2025年。 [在线]。获取地址：https: //openreview.net/forum?id=xmbdACI0xu [14] H. Touvron、L. Martin、K. Stone、P. Albert、A. Almahairi、Y. Babaei、N. Bashlykov、S. Batra、P. Bhargava、S. Bhosale 等，“Llama 2：开放基础模型与微调聊天模型”，《*arXiv* 预印本 *arXiv:2307.09288*》，2023年。[15] Qwen，：，A. Yang、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Li、D. Liu、F. Huang、H. Wei、H. Lin、J. Yang、J. Tu、J. Zhang、J. Yang、J. Yang、J. Zhou、J. Lin、K. Dang、K. Lu、K. Bao、K. Yang、L. Yu、M. Li、M. Xue、P. Zhang、Q. Zhu、R. Men、R. Lin、T. Li、T. Tang、T. Xia、X. Ren、X. Ren、Y. Fan、Y. Su、Y. Zhang、Y. Wan、Y. Liu、Z. Cui、Z. Zhang 和 Z. Qiu，“Qwen2.5 技术报告”，2025 [ ] 2412 15115 。在线。获取地址：https://arxiv.org/abs/。S. Antol、A. Agrawal、J. Lu、M. Mitchell、D. Batra、C. L. Zitnick 和 D. Parikh，“VQA：视觉问答”，载于 *IEEE* 国际计算机视觉会议论文集，2015年，第2425–2433页。

[17] Y. Li、Y. Du、K. Zhou、J. Wang、X. Zhao 和 J.-R. Wen，“评估大型视觉语言模型中的物体幻觉”，载于《2023年自然语言处理实证方法会议论文集》，H. Bouamor、J. Pino 和 K. Bali 编。新加坡：计算语言学协会，2023年12月，第292–305页。[在线]。获取地址：https://aclanthology.org/2023.emnlp-main.20/ [18] Y. Liu、H. Duan、Y. Zhang、B. Li、S. Zhang、W. Zhao、Y. Yuan、J. Wang、C. He、Z. Liu 等，“MMBench：你的多模态模型是全能选手吗？”，载于《欧洲计算机视觉会议》。施普林格，2024年，第216–233页。[19] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark 等，“通过自然语言监督学习可迁移视觉模型”，载于《国际机器学习会议》。PMLR，2021年，第8748–8763页。[20] Q. Sun、Y. Fang、L. Wu、X. Wang 和 Y. Cao，“EVA-CLIP：改进 CLIP 的大规模训练技术”，arXiv 预印本 arXiv:2303 15389，2023年。[21] Y. Wu、K. Chen、T. Zhang、Y. Hui、T. Berg-Kirkpatrick 和 S. Dubnov，“通过特征融合和关键词到描述增强进行大规模对比式语言-音频预训练”，载于《ICASSP 2023—2023年 IEEE 国际声学、语音与信号处理会议（ICASSP）》。IEEE，2023年，第1–5页。[22] L. Zheng、W.-L. Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang、Z. Lin、Z. Li、D. Li、E. Xing 等，“使用 MT-Bench 和 Chatbot Arena 评判 LLM-as-a-judge”，《神经信息处理系统进展》，第36卷，第46 595–46 623页，2023年。[23] J. Li、D. Li、S. Savarese 和 S. Hoi，“BLIP-2：借助冻结图像编码器和大语言模型实现语言-图像预训练”，载于《国际机器学习会议》。PMLR，2023年，第19 730–19 742页。[24] W. Dai、J. Li、D. Li、A. Tiong、J. Zhao、W. Wang、B. Li、P. N. Fung 和 S. Hoi，“InstructBLIP：迈向通过指令微调实现通用视觉语言模型”，《神经信息处理系统进展》，第36卷，第49 250–49 267页，2023年。[25] F. Li、R. Zhang、H. Zhang、Y. Zhang、B. Li、W. Li、Z. Ma 和 C. Li，“LLaVA-NeXT-Interleave：解决大型多模态模型中的多图像、视频和三维问题”，arXiv 预印本 arXiv:2407.07895，2024年。[26] H. Hu、Y. Zhou、L. You、H. Xu、Q. Wang、Z. Lian、F. R. Yu、F. Ma 和 L. Cui，“EmoBench-M：评估多模态大语言模型的情绪智力”，arXiv 预印本 arXiv:2502 04424，2025年。[27] Z. Lian、L. Sun、Y. Ren、H. Gu、H. Sun、L. Chen、B. Liu 和 J. Tao，“MERBench：多模态情绪识别的统一评估基准”，arXiv 预印本 arXiv:2401.03429，2024年。[28] Z. Lian、L. Sun、H. Sun、K. Chen、Z. Wen、H. Gu、B. Liu 和 J. Tao，“带有情绪能力的 GPT-4V：广义情绪识别的零样本基准”，《信息融合》，第108卷，文章编号102367，2024年。[在线]。获取地址：https://www.sciencedirect.com/science/article/pii/S1566253524001453[29] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过掩码预测隐藏单元进行自监督语音表征学习”，《IEEE/ACM音频、语音与语言处理汇刊》，第29卷，第3451–3460页，2021年。[30] A. v. d. Oord、Y. Li 和 O. Vinyals，“基于对比预测编码的表征学习”，arXiv 预印本 arXiv:1807.03748，2018年。[31] J. Devlin、M.-W. Chang、K. Lee 和 K. Toutanova，“BERT：用于语言理解的深度双向变换器预训练”，载于《2019年北美计算语言学协会年会：人类语言技术论文集》第1卷（长短篇论文），2019年，第4171–4186页。[32] F. Feng、Y. Yang、D. M. Cer、N. Arivazhagan 和 W. Wang，“语言无关的BERT句子嵌入”，载于《计算语言学协会年会》，2020年。 [在线]。获取地址：https://api.semanticscholar.org/CorpusID:220347683 [33] E. J. Hu、Y. Shen、P. Wallis、Z. Allen-Zhu、Y. Li、S. Wang、L. Wang、W. Chen 等，“LoRA：大语言模型的低秩适配”。ICLR，第1卷，第2期，第3页，2022年。[34] Y. Liu、W. Dai、C. Feng、W. Wang、G. Yin、J. Zeng 和 S. Shan，“MAFW：用于野外动态面部表情识别的大规模多模态复合情感数据库”，载于《第30届 ACM 多媒体国际会议论文集》，2022年，第24–32页。[35] X. Jiang、Y. Zong、W. Zheng、C. Tang、W. Xia、C. Lu 和 J. Liu，“DFEW：用于野外动态面部表情识别的大规模数据库”，载于《第28届 ACM 多媒体国际会议论文集》，2020年，第2881–2889页。[36] Z. Lian、R. Liu、K. Xu、B. Liu、X. Liu、Y. Zhang、X. Liu、Y. Li、Z. Cheng、H. Zuo 等，“Mer 2025：当情感计算遇上大型语言模型”，arXiv 预印本 arXiv:2504.19423，2025年。[37] Z. Jin、J. Jia、Q. Wang、K. Li、S. Zhou、S. Zhou、X. Qin 和 Z. Wu，“Speechcraft：一个带有自然语言描述的细粒度表达性语音数据集”，载于第32届 ACM 多媒体国际会议论文集，2024年，第1255–1264页。[38] DeepSeek-AI、D. Guo、D. Yang、H. Zhang、J. Song、R. Zhang、R. Xu、Q. Zhu、S. Ma、P. Wang、X. Bi、X. Zhang、X. Yu 和 Y. Wu，“Deepseek-r1：通过强化学习激发大语言模型的推理能力”，2025 [ ] 2501 12948 。在线。[39] 可访问：https://arxiv.org/abs/。DeepSeek-AI、A. Liu、A. Mei、B. Lin、B. Xue、B. Wang、B. Xu、B. Wu、B. Zhang、C. Lin、C. Dong、C. Lu、C. Zhao 和 C. Deng，“Deepseek-v3.2：推动开放式大语言模型的前沿”，2025 [ ] 2512 02556 。在线。可访问：[40] https://arxiv.org/。J. Ye、Z. Wang、H. Sun、K. Chandrasegaran、Z. Durante、C. Eyzaguirre、Y. Bisk、J. C. Niebles、E. Adeli、L. Fei-Fei、J. Wu 和 M. Li，“重新思考长视频理解中的时间搜索”，载于 IEEE/CVF 计算机视觉与模式识别会议（CVPR）论文集，2025年6月，第8579–8591页。[41] R. Brachman 和 H. Levesque，《知识表示与推理》。爱思唯尔，2004年。[42] D. Xu、W. Chen、W. Peng、C. Zhang、T. Xu、X. Zhao、X. Wu、Y. Zheng 和 E. Chen，“生成式信息抽取中的大型语言模型：综述”，《计算机科学前沿》，第 18， 186357, 卷。在线。2023 [ ] 可访问：https://api.semanticscholar.org/CorpusID: 266690657[43] Y. Chu、J. Xu、Q. Yang、H. Wei、X. Wei、Z. Guo、Y. Leng、Y. Lv、J. He、J. Lin 等，“Qwen2-Audio 技术报告”，arXiv 预印本 arXiv:2407.10759，2024年。[44] C. Tang、W. Yu、G. Sun、X. Chen、T. Tan、W. Li、L. Lu、Z. Ma 和 C. Zhang，“SALMONN：迈向大型语言模型的通用听觉能力”，载于第十二届国际学习表征会议（ICLR）论文集，2024年，扩展轨道。[45] S. Bai、K. Chen、X. Liu、J. Wang、W. Ge、S. Song、K. Dang、P. Wang、S. Wang、J. Tang 等，“Qwen2.5-VL 技术报告”，arXiv 预印本 arXiv:2502.13923，2025年。[46] Y. Li、C. Wang 和 J. Jia，“LLaMA-VID：一张图像在大型语言模型中价值两个词元”，载于《欧洲计算机视觉会议》。施普林格，2024年，第323–340页。[47] Q. Ye、H. Xu、G. Xu、J. Ye、M. Yan、Y. Zhou、J. Wang、A. Hu、P. Shi、Y. Shi 等，“mPLUG-Owl：模块化赋能大型语言模型实现多模态”，arXiv 预印本 arXiv:2304.14178，2023年。[48] M. Maaz、H. Rasheed、S. Khan 和 F. S. Khan，“Video-ChatGPT：借助大型视觉语言模型实现精细的视频理解”，arXiv 预印本 arXiv:2306.05424，2023年。[49] K. Li、Y. He、Y. Wang、Y. Li、W. Wang、P. Luo、Y. Wang、L. Wang 和 Y. Qiao，“VideoChat：以对话为中心的视频理解”，arXiv 预印本 arXiv:2305.06355，2023年。[50] K. Li、Y. Wang、Y. He、Y. Li、Y. Wang、Y. Liu、Z. Wang、J. Xu、G. Chen、P. Luo 等，“Mvbench：一个全面的多模态视频理解基准”，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2024年，第22 195–22 206页。[51] A. Yang、A. Li、B. Yang、B. Zhang、B. Hui、B. Zheng、B. Yu、C. Gao、C. Huang、C. Lv 等，“Qwen3 技术报告”，arXiv 预印本 arXiv:2505.09388，2025年。[52] K. Papineni、R. Roukos、T. Ward 和 W.-J. Zhu，“BLEU：一种机器翻译自动评估方法”，载于第40届计算语言学协会年会论文集，2002年，第311–318页。[53] R. Vedantam、C. Lawrence Zitnick 和 D. Parikh，“CIDEr：基于共识的图像描述评估”，载于 IEEE 计算机视觉与模式识别会议论文集，2015年，第4566–4575页。[54] B. Zhu、B. Lin、M. Ning、Y. Yan、J. Cui、W. HongFa、Y. Pang、W. Jiang、J. Zhang、Z. Li、C. W. Zhang、Z. Li、W. Liu 和 L. Yuan，“LanguageBind：通过基于语言的语义对齐，将视频语言预训练扩展至 N 模态”，载于第十二届国际学习表征会议，2024年。 [在线]。可访问：https://openreview.net/forum?id=QmZKc7UZCy [55] B. Zhang、P. Zhang、X. Dong、Y. Zang 和 J. Wang，“Long-clip：释放 CLIP 的长文本能力”，载于《欧洲计算机视觉会议》。施普林格，2024年，第310–325页。[56] J. Wang、C. Wang、K. Huang、J. Huang 和 L. Jin，“VideoCLIP-XL：提升视频片段模型的长描述理解能力”，载于2024年自然语言处理实证方法会议（EMNLP）论文集。佛罗里达州迈阿密：计算语言学协会