arXiv:2603.14976v1 [cs.MM] 16 Mar 2026

文本中的情绪锚定：用于模仿强度估计的鲁棒多模态融合

Lingsi Zhu[1] Yuefeng Zou[1] Yunxiang Zhang[1] Naixiang Zheng[1] Guoyuan Wang[1] JunYu[1]*Jiaen Liang[2] Wei Huang[2] ShengpingLiu[2] XiminZheng[3]1中国科学技术大学 [2]云知声智能科技股份有限公司[3]平安科技有限公司[1]{ls-zhu24, zouyuefeng.00, mesa, zhengnx, wgy2874849700}@mail.ustc.edu.cn[1] harryjun@ustc.edu.cn [2]{liangjiaen, huangwei, liushengping}@unisound.com[3]ZHENGXIMIN135@pingan.com.cn

摘要

在自然环境中估计情绪模仿强度（*EMI*）是情感计算领域一项至关重要却颇具挑战性的任务。其主要难点在于，如何有效建模高度异构模态之间复杂且非线性的时间动态，尤其是在物理信号受到干扰或缺失时。为解决这一问题，我们提出 *TAEMI*（文本锚定情绪模仿强度估计），这是一种专为第十届 *ABAW* 竞赛设计的新型多模态框架。鉴于连续视觉和声学信号极易受到瞬态环境噪声影响，我们打破了传统的对称融合范式，转而利用文本转录——其天然编码了稳定且与时间无关的语义先验——作为核心锚点。具体而言，我们提出一种文本锚定双重交叉注意力机制，利用这些鲁棒的文本查询主动滤除帧级冗余，并对齐受噪声干扰的物理信号流。此外，为避免在不受约束的真实场景中因不可避免的数据缺失而导致性能灾难性退化，我们在训练期间引入可学习缺失模态令牌和模态丢弃策略。在 *Hume-Vidmimic2* 数据集上的大量实验表明，*TAEMI* 能够有效捕捉细粒度的情绪变化，并在条件不完善时保持稳健的预测能力。我们的框架在六个连续情绪维度上的平均皮尔逊相关系数达到当前最先进水平，显著优于现有基线方法。

1. 引言

情绪模仿是人类社会互动中的一种基本机制，个体会在无意识中复制他人的面部表情、语音韵律和身体动作，以建立共情并加强社会纽带。赋予机器准确感知和量化这些动态情绪变化的能力，已成为情感计算中的一项关键挑战。因此，情绪模仿强度（EMI）的自动估计——一项旨在捕捉情绪传递中连续且细微变化的任务——已成为一个前沿研究方向，对虚拟助手、社交机器人学和心理健康评估具有深远意义。

然而，现有方法在应对野外环境下的 EMI 估计任务时面临着显著瓶颈。首先，情绪模仿本质上是一个复杂的多模态时间过程。细微的面部表情与动态的声音波动会随时间发生非线性交织。当前研究主要依赖单模态分析或简单的后期融合策略（例如简单特征拼接），导致模态间协同作用未得到充分利用。例如，采用 ViT[2] 提取视觉特征的基线模型，其平均皮尔逊相关系数仅为 0.09，这表明单模态视觉网络难以捕捉情绪强度的时间演变。其次，细粒度的时间对齐需要强大的跨模态语义理解能力。虽然 Wav2Vec2 [1] 等声学特征提取方法取得了相对较高的 0.24 相关系数，但它们缺乏必要的语义基础，难以在嘈杂且不受约束的环境中保持稳健性。现有融合框架将所有模态一视同仁，未能建立稳定的语义锚点来引导异构情绪线索的对齐。

*通讯作者。

为应对这些关键挑战，我们提出 TAEMI（文本锚定情绪模仿强度估计），这是一种专为第十届 ABAW（野外环境下的情感行为分析）挑战赛 EMI赛道设计的新型多模态框架 [5–24, 36]。TAEMI 的架构设计由两个核心动机驱动：

1）通过语义锚点打破融合对称性：我们发现，连续的高帧率视觉与声学信号极易受到瞬时环境噪声和时间错位的影响。相比之下，文本转录明确编码了稳定且不受时间影响的语义上下文。因此，TAEMI 不再被动拼接特征，也不再对所有模态一视同仁，而是引入文本锚定的双重交叉注意力机制。该机制以全局文本表征作为鲁棒的查询锚点，主动滤除帧级冗余，并使高度动态的视觉与声学键围绕稳定的语义核心对齐。

2）防止真实环境中的灾难性退化：在不受约束的真实环境中，情感信号经常受到干扰，甚至完全缺失（例如，由面部遮挡或麦克风故障所致）。严格依赖所有模态同时存在的模型，在这种不完美条件下会遭遇灾难性的性能退化。为缓解这一脆弱性，我们整合了两种协同作用的鲁棒性机制。可学习缺失模态令牌提供明确的占位表征，在物理数据缺失时维持网络稳定；而在训练期间采用的模态丢弃策略，则迫使网络动态平衡其预测能力，避免过度依赖任何单一模态。

我们在 Hume-Vidmimic2 数据集上进行了系统性实验，该数据集包含 15,000 个视频，总时长超过 25 小时，呈现了复杂的模仿互动。实验结果验证了我们方法的优越性。

总而言之，本研究的主要贡献有三点：

- 我们提出了 TAEMI，这是一种新颖的多模态架构，将融合范式从简单拼接转变为文本锚定双重交叉注意力机制，利用文本作为语义指引，有效对齐视觉和声学时间线索。
- 我们提出了以鲁棒性为导向的训练策略，包括可学习缺失模态令牌和模态丢弃，显著增强了模型应对不完整真实世界情感数据的容错能力。
- 我们的框架在 Hume-Vidmimic2 数据集上创下新的最先进水平，在六个情绪维度上的平均皮尔逊相关系数达到 0.55，大幅超越官方基线和现有方案。

2. 相关工作

2.1. 多模态情绪识别技术

多模态情绪识别从根本上依赖于融合视觉、声学和文本通道中的互补信息，以增强情感分析的鲁棒性。张量融合网络（TFN） [35] 和动态融合网络（DFN） [32], 等早期融合范式引入了模态间门控机制。然而，静态权重分配限制了它们适应情绪强度流动的时间动态的能力。为了解决动态交互问题，后续研究转向了先进的跨模态建模。例如，多模态 Transformer [29] 能够在不同时间步将一个模态的流隐式适配到另一个模态。与此同时，MISA [4]将不同模态映射到彼此独立的不变子空间和特定子空间，以减少模态间差异，并在多种情感分析任务中展现出出色的泛化能力。

尽管这些方法能够有效建模一般性的跨模态交互，但它们通常对所有模态一视同仁。在处理高度异质的数据时，这种对称处理可能会加剧语义错位。相比这些现有方法，我们的 TAEMI 框架明确打破了这种对称性。通过将文本模态——其本身承载着高层次的语义先验——指定为核心锚点，我们主动引导更抽象的视觉和声学信号进行时间对齐与融合，从而减轻由模态异质性所造成的语义差异。

2.2. 情感计算与情绪模仿强度

情感计算领域已逐步从静态的类别化情绪分类，转向对现实环境中人类情感行为进行连续的多维回归。在这一背景下，估计情绪模仿强度（EMI）——量化个体在社交互动过程中如何无意识地模仿他人——已成为一项重要任务。

早期的 EMI [3] 方法主要依赖单模态特征，例如 Wav2Vec 2.0 声学表征或视觉 Transformer 视觉描述符。然而，近期最先进水平的方法 [28, 33, 34, 38] 强调，稳健的多模态整合绝对不可或缺。研究人员认识到，物理声学和视觉信号极易受到环境噪声影响，且单独使用时跨模态互补性较弱，因此日益将文本模态纳入其中。近期研究表明，从转录语音中提取的语义上下文（例如通过 Whisper [27]获得）可作为主要先验，显著助力复杂情绪状态的解读。

为弥合异构模态之间的表征差距，现代框架采用对比式架构（例如基于 CLIP 的对齐[26]）进行模态解耦预训练。此外，研究人员提出了利用时序卷积网络（TCN[25]）和长短期记忆网络的动态融合网 [37] 络，以自适应地分配注意力权重，力求提升在开放世界环境中应对遮挡和噪声的稳健性。尽管取得了这些重大进展，连续物理信号与离散语义词元之间细粒度的时间错位仍是一个关键瓶颈。我们的工作通过引入文本锚定融合流程，并配合显式的缺失模态稳健性机制，直接解决这一问题，确保充分发挥这些模态的互补特性并稳健地加以维持。

3. 方法

3.1. 问题表述

令 $x_i = \{x_i^a, x_i^v, x_i^t\}$ 表示第 i 个多模态输入样本，其中上标 a, v 和 t 分别表示声学、视觉和文本模态。我们的目标是在实坐标空间 $\mathbb{R}^6$中预测一个 6 维连续情绪强度向量 $\hat{\mathbf{y}}_i \in \mathbb{R}^6$，用于量化六种目标情绪的强度：钦佩、愉悦、决心、共情痛苦、兴奋和喜悦。给定相应的真实强度向量 $\mathbf{y}_i \in \mathbb{R}^6$，我们将此任务表述为多标签回归任务，并通过最小化预测情绪强度与真实情绪强度之间的均方误差（MSE）来优化模型。

3.2. 数据处理与序列构建

为了为每个多模态样本构建时间对齐的表示，我们的输入流程按以下步骤处理数据：

音频： 我们将原始声学波形截断或补零至固定的 12 秒时长。随后，Wav2Vec 2.0 [1] 特征提取器处理这些波形，生成注意力掩码，以区分有效音频帧与填充区域。

视觉： 我们使用预训练的 ViT [2, 30]模型提取视觉帧特征，并将时间序列截断或填充至最多 400 帧。

文本： 我们采用 Whisper [27] 模型转录音频。随后，通过 GTE[39]分词器对转录词元进行编码，并将最大序列长度严格限制为 128 个词元。

此外，我们明确保留缺失模态的指示信息。在优化过程中，我们利用注意力掩码在批次内整理长度不一的序列，从而准确跳过填充位置。

3.3. 文本锚定双重交叉注意力融合

如图 1所示，我们的架构通过精心设计的文本锚定注意力流程提取并融合特征。我们首先定义每种模态的初始表示。

音频表示： 我们将组合后的音频矩阵定义为：

$$\mathbf{A} = \mathrm{Concat}\left(\mathbf{H}^a, \mathbf{Z}^a\right) \in \mathbb{R}^{T_a \times 1027},$$

其中，Concat(·)表示拼接操作，$\mathbf{H}^a$表示隐藏状态序列，$\mathbf{Z}^a$ 表示从预训练情感音频模型中提取的辅助逻辑值，T_a 表示音频模态的时间序列长度。

文本表示： 我们提取文本特征向量如下：

$$\mathbf{t} = \mathrm{CLS}\big(\mathrm{GTE}(x^t)\big) \in \mathbb{R}^{768},$$

其中，GTE(·) 表示文本编码器，CLS(·) 提取全局分类标记。

视觉表示： 我们将视觉输入表示为$\mathbf{V} \in \mathbb{R}^{T_v \times 768}$，其中 T_v 表示视觉帧的时间序列长度。

为对齐这些异构模态，我们首先使用简单的线性层，将文本、音频和视觉特征投影到共享潜在空间中。完成模态对齐后，我们将每种模态的序列分别输入专用的自注意力层，以捕捉模态内的时间依赖关系。

随后，对齐并经时间建模的表示（分别记为音频和视觉的 $\mathbf{A}'$ 和 $\mathbf{V}'$）进入双重交叉注意力层。将文本模态设为中心锚点，主要是因为它具有时间不变性和丰富的语义信息。与连续、高帧率且极易受到瞬时物理噪声影响的视觉和声学信号不同，全局文本表征 $\mathbf{t}$ 是一种稳定、独立于时间的语义先验。因此，通过使用单向量文本查询$\mathbf{t}$ 关注非文本模态中含噪的时间键和值，我们的架构能够有效滤除帧级冗余，并围绕稳健的语义核心对齐异构流：

$$\mathbf{h}^a = \mathrm{CA}_a(\mathbf{t}, \mathbf{A}'), \quad \mathbf{h}^v = \mathrm{CA}_v(\mathbf{t}, \mathbf{V}'),$$

其中，$\mathrm{CA}_a(\cdot)$ 和 $\mathrm{CA}_v(\cdot)$ 表示特定于模态的交叉注意力操作，为$\mathbf{h}^a$ 和$\mathbf{h}^v$生成 512 维输出向量。

我们通过拼接各模态专属向量来构建融合后的多模态表示：$\mathbf{f} = [\mathbf{t}; \mathbf{h}^a; \mathbf{h}^v] \in \mathbb{R}^{1792}$，其中 $[\cdot;\ \cdot]$ 表示沿特征维度进行向量拼接。最后，一个多层

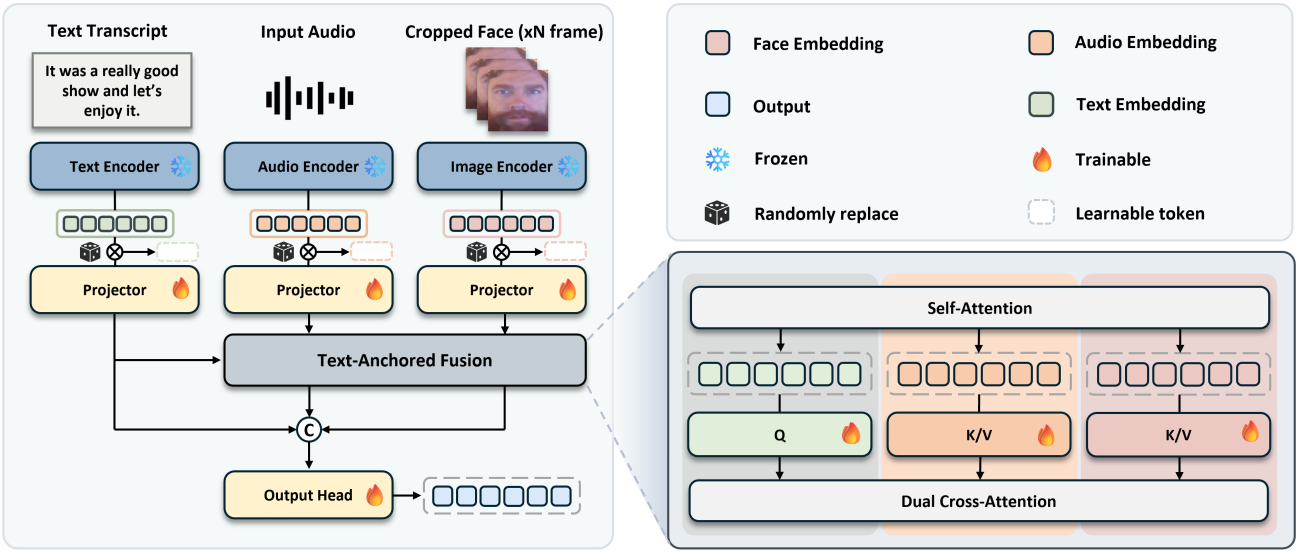


图1. **TAEMI 的整体架构**。该框架处理经过时间对齐的音频、视觉和文本模态。在进行模态特定的特征提取和线性对齐后，时间表示被输入文本锚定双重交叉注意力模块。在此，文本特征作为查询，用于汇聚相关的音频和视觉上下文。随后，多层感知机（MLP）对融合后的多模态表示进行处理，以预测连续情绪强度向量。在训练期间，策略性地引入可学习缺失模态令牌和模态丢弃，以增强模型的鲁棒性。

感知机（多层感知机）头对该融合向量进行处理，生成最终预测：

$$\hat{\mathbf{y}}_i = \mathbf{W}_2 \, \mathrm{Dropout}(\mathrm{GELU}(\mathbf{W}_1 \mathbf{f})) \, ,$$

其中，$\mathbf{W}_1$ 和$\mathbf{W}_2$ 是多层感知机的可学习权重矩阵，GELU(·)是高斯误差线性单元激活函数，而 Dropout(·) 是丢弃正则化操作。

3.4. 鲁棒性机制

为增强模型在真实场景中应对缺失或噪声数据的能力，我们在架构中融入了两种旨在提升鲁棒性的机制：

可学习缺失模态令牌：当输入样本中明确缺少视觉或文本模态时，我们会注入对应模态专属的可训练占位令牌。

模态丢弃：在训练期间，我们以 $p = 0.1$ 的概率随机用各模态对应的缺失令牌替换该模态。此策略促使网络避免过度依赖任何单一模态。

3.5. 优化策略

我们使用 AdamW 优化器对网络进行优化，学习率为 1×10^{-4}，权重衰减为 0.01。我们采用余弦学习率调度，并进行一轮线性预热。训练采用混合精度、批量大小为 32，并采用耐心值为 5 个训练轮次的早停策略。目标函数为均方误差（MSE）损失，定义如下：

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{6N} \sum_{i=1}^{N} \|\hat{\mathbf{y}}_i - \mathbf{y}_i\|_2^2 \, ,$$

其中，$\mathcal{L}_{\mathrm{MSE}}$ 表示损失值，N 表示一个批次中的样本总数，$\|\cdot\|_2^2$ 表示平方 L_2 范数。

4. 实验

4.1. 设置

本节详细介绍我们的实验设置，包括数据集、评估指标和优化流程。

数据集。我们在多模态情感数据集 Hume-Vidmimic 上进行了所有实验2 [21–24]。该数据集包含 15,000 个视频样本，涵盖六个连续的情绪维度：钦佩、愉悦、决心、共情痛苦、兴奋和喜悦。专业标注团队按照标准化流程，对每个样本的情绪强度进行标注，分数范围为 $[0, 1]$ 。表1 展示了训练集、验证集和测试集的统计详情。

评估。我们采用平均皮尔逊相关系数（ρ）作为强度估计的主要评估指标。该指标能够稳健地衡量预测的情绪反应强度与真实值之间的线性相关性——

表 1. **Hume-Vidmimic2 数据集各划分的统计概览**。该数据集经过严格划分，以确保训练、验证和测试阶段之间不存在数据泄漏。

划分	时长	样本
训练	15:07:03	8072
验证	9:12:02	4588
Test	9:04:05	4586
$\sum$	33:23:10	17246

我们将此指标定义为：

$$\rho = \frac{1}{6}\sum_{i=1}^{6}\rho_i$$

其中，$\rho_i (i \in \{1, 2, \ldots, 6\})$ 表示六种情绪类别中每一种的相关系数，其定义如下：

$$\rho_i = \frac{\mathrm{cov}(y_i, \hat{y_i})}{\sqrt{\mathrm{var}(y_i)\mathrm{var}(\hat{y_i})}}$$

这里，$\mathrm{cov}(y_i, \hat{y}_i)$ 表示预测值与目标值之间的协方差，而 $\mathrm{var}(y_i)$ 和 $\mathrm{var}(\hat{y}_i)$ 分别表示它们各自的方差。

训练细节。为确保比较公平，我们在受控条件下进行所有消融实验，保留基线优化方案，并在每次运行中仅修改一个架构因素。我们最多训练模型 50 轮，并采用耐心值为 5 轮的早停策略。我们使用 AdamW 优化器（权重衰减 $= 0.01$）优化网络，初始学习率为 1×10^{-4}。学习率调度器采用线性预热（1 轮）结合余弦衰减，并在每个优化步骤进行更新。为加快训练，我们启用混合精度，通过梯度缩放进行训练，并采用阈值为 100的梯度裁剪。我们将批量大小设为 32，梯度累积步数设为 1，模态丢弃概率设为 0.1。

对于多模态骨干网络的初始化，我们从 Wav2Vec 2.加载音频网络。0 [1, 31], 使用 GTE 初始化文本编码器 [39], 并直接使用预提取的 ViT 表征作为视觉特征。所有实验均在单块 NVIDIA A800-SXM4-80GB GPU 上运行，并在每个训练轮次监测验证集皮尔逊相关系数（ρ_{val}）。

4.2. 主要结果

表 2 总结了我们提出的 TAEMI 架构与现有先进解决方案在 EMI 估计挑战赛验证集上的性能对比。结果表明，仅依赖单模态或双模态输入会不可避免地限制预测能力。通过文本锚定融合机制有效整合全部三种模态（$V + A + T$），TAEMI 的平均 ρ_{val} 达到**0.55**，显著优于基线和先前的顶尖解决方案 [34]。

4.3. 消融研究

为验证 TAEMI 中核心架构组件的各自贡献，我们开展了全面的消融研究。具体而言，我们系统地分离并评估了两项关键机制：（1）文本锚定双重交叉注意力捕捉跨模态动态的能力；（2）我们的鲁棒性模块在提升模型对噪声或缺失数据耐受性方面的有效性。表3详细列出了这些独立变体的定量结果。

文本锚定融合的有效性。为对我们的融合策略建立严格的下界，我们首先评估了一个简单特征拼接基线（表3，第 1 行）。在该配置中，我们完全绕过注意力流程——移除了模态内自注意力和文本锚定交叉注意力模块。相反，视觉、声学和文本模态的线性投影表示被简单地展平，并沿特征维度拼接后直接输入最终的多层感知机头。如表3所示，这种朴素的融合方法仅取得了 0.45 的次优性能。此外，仅使用模态内自注意力而不采用连接它们的文本锚定交叉注意力，会使性能受限于0.48。这一递进关系清晰地验证了我们的核心动机：由于文本明确编码了人类表达的潜在语义上下文，作为稳定锚点（查询）来高效聚合相关视觉和声学线索（键/值），能够对齐异质的情感信号，从而相较于自注意力变体带来显著的 +0.07性能提升。

鲁棒性机制的影响。现实世界中的情感数据经常存在模态缺失。移除显式的可学习缺失令牌后，性能降至 0.52，这表明当某个模态缺失时，标准零填充无法为网络提供足够的语义基础。更重要的是，在训练期间禁用模态丢弃（$p = 0$）会导致性能明显下降至 0.51。这证实了若不采用模态丢弃，网络往往会对所有模态同时存在的情况过拟合，在不完美的验证样本上进行评估时便会失去预测韧性。整合这两种机制后，TAEMI 得以达到峰值性能。

锚定模态的选择。 为严谨地论证我们选择文本作为中心查询锚点的架构设计，我们评估了其他模态排列方式。具体而言，我们构建了一个视觉锚定 变体（其中视觉特征作为查询，用于关注音频

表 2. **EMI 估计挑战赛验证集上的性能比较**。我们报告六个情感目标的平均皮尔逊相关系数（ρ_{val}）。 V、 A 和 T 分别表示视觉、声学和文本模态。最佳结果以 **粗体**标出。

类别	模型	模态	平均值 ρ_{val}
基线	ViT（基线） [19]	V	0.09
	Wav2Vec2（基线） [19]	A	0.24
	ViT +Wav2Vec2（基线） [19]	$V+A$	0.25
先前解决方案	Yu等人 [33]	$V+A$	0.32
	Hallmen等人 [3]	A	0.38
	Savchenko等人 [28]	$V+A+T$	0.44
	Zhang等人 [38]	$V+A+T$	0.46
	Yu等人 [34]	$V+A+T$	0.51
Ours	**TAEMI**	$V+A+T$	**0.55**

表3。**TAEMI关键组件的消融研究**。 我们采用留一评估，以分离文本锚定交叉注意力（TA-CA）、可学习缺失令牌（ LMT）和模态丢弃（MD）各自的贡献。 Δ 列明确展示了移除特定模块后性能的下降。

TA-CA	LMT	MD	均值 ρ_{val} (↑)	Δ
简单特征拼接			0.45	-0.10
×	✓	✓	0.48	-0.07
✓	×	✓	0.52	-0.03
✓	✓	×	0.51	-0.04
✓	✓	✓	**0.55**	-

表4。**双重交叉注意力机制中不同锚定模态的影响**。 我们比较了将视觉、音频或文本作为中心查询，以关注另外两种模态的效果。由于文本具有较高的语义稳定性，因此它是最有效的查询锚点。

锚定模态（查询）	均值 ρ_{val} (↑)
视觉锚定	0.39
音频锚定	0.51
文本锚定	**0.55**

和文本的键/值）以及一个音频锚定 变体（其中音频特征作为查询，用于关注视觉和文本）。表4 显示，视觉锚定和音频锚定模型的平均皮尔逊相关系数分别为 0.49 和 0.51，均显著低于我们采用的文本锚定方案（0.55）。

这种性能差距与我们对多模态情感数据的物理直觉高度吻合。视觉和声学信号是连续的、帧率高的，而且极易受到物理环境噪声的影响，例如瞬时遮挡、光照变化或背景声音重叠。因此，将它们用作查询往往会导致注意力图不稳定。相比之下，文本转录内容代表离散的高层语义先验，能够天然地抽象掉低层物理噪声。文本作为查询锚点，可提供稳健且不受噪声影响的语义坐标，在不被帧级波动干扰的情况下，成功引导冗余物理信号的时间对齐。

5. 结论

本文提出了 TAEMI，这是一种稳健且高效的多模态框架，用于连续估计情绪模仿强度（EMI）。我们发现，对称晚期融合方法无法弥合高度异质的情感信号之间的语义鸿沟。为解决这一问题，我们通过引入文本锚定双重交叉注意力机制，改变了融合范式。通过明确将文本模态指定为语义锚点，我们的架构成功引导视觉和声学特征进行细粒度的时间对齐。此外，我们证明，将可学习缺失模态令牌与模态丢弃训练策略结合使用，可显著增强网络应对无约束环境中常见的模态缺失或噪声数据的能力。在第十届 ABAW 竞赛的挑战性 Hume-Vidmimic2 数据集上进行评估后，TAEMI 以 0.55 的总体平均皮尔逊相关系数创下最先进水平。未来，我们计划探索集成大语言模型（ LLMs），以提取更深层次的语义先验，进一步提升多模态情感计算在真实环境中的能力。

6. 致谢

本研究得到中国自然科学基金（62276242）、合肥市自然科学基金（HZR2431）以及基于 OpenI社区开发的 CAAI-MindSpore开放基金的资助。

参考文献

[1] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。Wav2Vec 2.0：语音表征自监督学习框架，2020年。1、3、5[2] Alexey Dosovitskiy、Lucas Beyer、Alexander Kolesnikov、Dirk Weissenborn、Xiaohua Zhai、Thomas Unterthiner、Mostafa Dehghani、Matthias Minderer、Georg Heigold、Sylvain Gelly、Jakob Uszkoreit 和 Neil Houlsby。一张图像胜过 16×16 个单词：用于大规模图像识别的 Transformer，2021年。1、3[3] Tobias Hallmen、Fabian Deuser、Norbert Oswald 和 Elisabeth André。用于情绪模仿强度预测的单模态多任务融合，2024年。2、6[4] Devamanyu Hazarika、Roger Zimmermann 和 Soujanya Poria。MISA：用于多模态情感分析的模态不变与模态特异表征。载于第28届ACM国际多媒体会议论文集，页码 1122 1131， 2020 2 –，美国华盛顿州西雅图，。[5] Dimitrios Kollias。ABAW：效价-唤醒度估计、表情识别、动作单元检测与多任务学习挑战赛。载于IEEE/CVF计算机视觉与模式识别会议论文集，页码2328–2336，2022年。2[6] Dimitrios Kollias。ABAW：从合成数据中学习与多任务学习挑战赛。载于欧洲计算机视觉会议，页码157–172。施普林格，2023年。[7] Dimitrios Kollias。多标签复合表情识别：C-expr数据库与网络。载于IEEE/CVF计算机视觉与模式识别会议论文集，页码5589–5598，2023年。[8] Dimitrios Kollias 和 Stefanos Zafeiriou。表情、情感与动作单元识别：Aff-Wild2、多任务学习和ArcFace。arXiv预印本 arXiv:1910.04855，2019年。[9] Dimitrios Kollias 和 Stefanos Zafeiriou。野外场景情感分析：效价-唤醒度、表情、动作单元与统一框架。arXiv预印本 arXiv:2103.15792，2021年。[10] Dimitrios Kollias 和 Stefanos Zafeiriou。第二届ABAW2竞赛中的情感行为分析。载于IEEE/CVF国际计算机视觉会议论文集，页码3652–3660，2021年。[11] D Kollias、A Schulc、E Hajiyev 和 S Zafeiriou。首届ABAW 2020竞赛中的情感行为分析。载于2020年第15届IEEE国际自动人脸与手势识别会议（FG 2020）（FG），页码794–800，。[12] Dimitrios Kollias、Panagiotis Tzirakis、Alan Cowen、Irene Kotsia、UK Cogitat、Eric Granger、Marco Pedersoli、Simon Bacon、Alice Baird、Chunchang Shao 等。情感与行为分析的进展：第八届ABAW研讨会与竞赛。.

[13]Dimitrios Kollias、Viktoriia Sharmanska 和 Stefanos Zafeiriou。多合一面部行为分析：在单一网络中处理表情、情感和动作单元。arXiv预印本 arXiv:1910.11111，2019年。[14] Dimitrios Kollias、Panagiotis Tzirakis、Mihalis A Nicolaou、Athanasios Papaioannou、Guoying Zhao、Bj¨orn Schuller、Irene Kotsia 和 Stefanos Zafeiriou。野外环境中的深度情感预测：Aff-wild数据库与挑战赛、深度架构及其他进展。《国际计算机视觉杂志》，第1–23页，2019年。[15] Dimitrios Kollias、Viktoriia Sharmanska 和 Stefanos Zafeiriou。异构多任务学习的分布匹配：一项大规模人脸研究。arXiv预印本 arXiv:2105.03790，2021年。[16] Dimitrios Kollias、Panagiotis Tzirakis、Alice Baird、Alan Cowen 和 Stefanos Zafeiriou。ABAW：效价-唤醒度估计、表情识别、动作单元检测与情绪反应强度估计挑战赛。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第5888–5897页，2023年。[17] Dimitrios Kollias、Chunchang Shao、Odysseus Kaloidas 和 Ioannis Patras。Behaviour4All：野外环境下面部行为分析工具包。arXiv预印本 arXiv:2409.17717，2024年。[18]Dimitrios Kollias、Viktoriia Sharmanska 和 Stefanos Zafeiriou。分类任务多任务学习的分布匹配：在人脸及其他领域开展的大规模研究。载于《AAAI人工智能会议论文集》，第2813–2821页，2024年。[19] Dimitrios Kollias、Panagiotis Tzirakis、Alan Cowen、Stefanos Zafeiriou、Irene Kotsia、Alice Baird、Chris Gagne、Chunchang Shao 和 Guanyu Hu。第六届野外情感行为分析竞赛（ABAW）。载于《IEEE/CVF计算机视觉与模式识别会议论文集》，第4587–4598页，2024年。6[20] Dimitrios Kollias、Stefanos Zafeiriou、Irene Kotsia、Abhinav Dhall、Shreya Ghosh、Chunchang Shao 和 Guanyu Hu。第七届ABAW竞赛：多任务学习与复合表情识别。arXiv预印本 arXiv:2407 03835，2024年。.[21]Dimitrios Kollias、Damith C Senadeera、Jianian Zheng、Kaushal KK Yadav、Greg Slabaugh、Muhammad Awais 和 Xiaoyun Yang。DVD：用于推进真实场景中暴力检测的综合数据集。arXiv预印本 arXiv:2506.05372，2025年。4[22] Dimitrios Kollias、Panagiotis Tzirakis、Alan Cowen、Stefanos Zafeiriou、Irene Kotsia、Eric Granger、Marco Pedersoli、Simon Bacon、Alice Baird、Chris Gagne 等。情感与行为分析的进展：第八届ABAW研讨会与竞赛。载于《计算机视觉与模式识别会议论文集》，第5572–5583页，2025年。[23] Dimitrios Kollias、Panagiotis Tzirakis、Alan S. Cowen、Stefanos Zafeiriou、Irene Kotsia、Eric Granger、Marco Pedersoli、Simon L. Bacon、Alice Baird、Chris Gagne、Chunchang Shao、Guanyu Hu、Soufiane Belharbi 和 Muhammad Haseeb Aslam。情感与行为分析的进展：第八届ABAW研讨会与竞赛。2025年。

[24] Dimitrios Kollias、Stefanos Zafeiriou、Irene Kotsia、Greg Slabaugh、Damith Chamalke Senadeera、Jianian Zheng、Kaushal Kumar Keshlal Yadav、Chunchang Shao 和 Guanyu Hu。从情绪到暴力：第9届ABAW中的多模态细粒度行为分析。载于《IEEE/CVF国际计算机视觉会议论文集》，第1–12页，2025年。2，4[25] Colin Lea、Michael D. Flynn、Rene Vidal、Austin Reiter 和 Gregory D. Hager。用于动作分割与检测的时序卷积网络，2016年。3[26]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark、Gretchen Krueger 和 Ilya Sutskever。通过自然语言监督学习可迁移视觉模型，2021年。3[27] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey 和 Ilya Sutskever。通过大规模弱监督实现鲁棒语音识别，2022年。2，3[28] Andrey V. Savchenko。Hsemotion团队参加ABAW-8竞赛：视听矛盾/犹豫、情绪模仿强度与面部表情识别，2025年。2，6[29] Yao-Hung Hubert Tsai、Shaojie Bai、Paul Pu Liang、J. Zico Kolter、Louis-Philippe Morency 和 Ruslan Salakhutdinov。用于未对齐多模态语言序列的多模态Transformer，2019年。2[30] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 Illia Polosukhin。Attention Is All You Need，2023年。3[31] Johannes Wagner、Andreas Triantafyllopoulos、Hagen Wierstorf、Maximilian Schmitt、Felix Burkhardt、Florian Eyben 和 Bj¨orn W. Schuller。语音情绪识别Transformer时代的曙光：弥合效价差距。《IEEE模式分析与机器智能汇刊》，45(9)：10745–10759，2023年。5[32] Yichong Xu、Jingjing Liu、Jianfeng Gao、Yelong Shen 和 Xiaodong Liu。用于机器阅读理解的动态融合网络，2018年。2[33] Jun Yu、Wangyuan Zhu 和 Jichao Zhu。视听情绪模仿强度估计的高效特征提取与后期融合策略，2024年。2，6[34] Jun Yu、Lingsi Zhu、Yanjun Chi、Yunxiang Zhang、Yang Zheng、Yongqi Wang 和 Xilong Lu。第八届野外情感行为分析竞赛中EMI挑战赛的技术方案，2025年。2，5，6[35] Amir Zadeh、Minghai Chen、Soujanya Poria、Erik Cambria 和 Louis-Philippe Morency。用于多模态情感分析的张量融合网络，2017年。2[36] Stefanos Zafeiriou、Dimitrios Kollias、Mihalis A Nicolaou、Athanasios Papaioannou、Guoying Zhao 和 Irene Kotsia。Aff-wild：野外环境中的效价与唤醒度挑战赛。载于《计算机视觉与模式识别研讨会》（CVPRW），2017年IEEE会议论文集，第1980–1987页。IEEE，2017年。2[37] Shu Zhang、Dequan Zheng、Xinchen Hu 和 Ming Yang。用于关系分类的双向长短期记忆网络。载于《第29届太平洋亚洲语言、信息与计算会议，第73–78页，中国上海，2015年。3[38] Wei Zhang、Feng Qiu、Chen Liu、Lincheng Li、Heming Du、Tiancheng Guo 和 Xin Yu。通过整合多模态知识进行情感行为分析，2024年。2，6[39] Xin Zhang、Yanzhao Zhang、Dingkun Long、Wen Xie、Ziqi Dai、Jialong Tang、Huan Lin、Baosong Yang、Pengjun Xie、Fei Huang、Meishan Zhang、Wenjie Li 和 Min Zhang。mgte：用于多语言文本检索的通用长上下文文本表示与重排序模型，2024年。3，5