arXiv:2603.10043v1 [cs.MM] 7 Mar 2026

AMB-DSGDN：面向多模态情感识别的自适应模态平衡动态语义图差异网络

YunshengWang[a]，YuntaoShou[a]，YilongTan[a]，Wei Ai[a]，Tao Meng[a]，*和Keqin Li[b]

[a]机构=中南林业科技大学计算机与数学学院，邮编=410004，城市= 湖南省长沙市，国家=中国

[b]机构=计算机科学系，纽约州立大学，城市=纽约州新帕尔茨，邮编=12561，国家=美国

A

文献信息

关键词：多模态情感识别 自适应模态平衡 说话人语义图 图神经网络

摘要

多模态对话情感识别通过融合文本、视觉和音频模态来捕捉情感线索。然而，现有方法在情感依赖关系建模和多模态表示学习方面仍存在明显局限。一方面，这些方法无法有效滤除多模态特征中的冗余或噪声信号，从而妨碍准确捕捉说话人之间及说话人内部情感状态的动态变化。另一方面，在多模态特征学习过程中，主导模态（如文本线索）往往会在融合过程中占据主导地位，进而抑制语音和视觉等非主导模态的互补作用，最终限制整体识别性能。为应对这些挑战，我们提出一种自适应模态平衡动态语义图差分网络（AMB-DSGDN）。具体而言，我们首先为文本、语音和视觉构建模态特定子图，其中每种模态均包含说话人内部图和说话人间图，以捕捉自身连续性和跨说话人情感依赖。在此基础上，我们引入一种差分图注意力机制，用于计算两组注意力图之间的差异。通过显式对比这些注意力分布，该机制能够抵消共享噪声模式，同时保留模态特定和上下文相关信号，从而获得更纯净、更具判别性的情感表示。此外，我们设计了一种自适应模态平衡机制，根据各模态对情感建模的相对贡献，为每种模态估算丢弃概率。该机制会随机丢弃部分特征

占主导地位的模态以抑制其过度影响，同时根据丢弃概率按比例重新缩放保留的特征，以维持整体信息平衡。在 IEMOCAP 和 MELD 数据集上的大量实验验证了 AMB-DSGDN 显著优于当前最先进的基线，证明了其在多模态对话情绪识别中的有效性和鲁棒性。

1. 引言

对话情绪识别是人机交互、自然语言处理和情感计算中的一项关键任务，旨在准确识别多方对话中说话人的情感状态，从而提升智能系统在社交机器人、虚拟助手、心理健康监测和客户服务等应用中的性能 [1], [2], [3], [4]。随着人工智能的快速发展，对话情绪识别已从单模态方法发展为多模态方法，整合文本、视觉和音频来源，以捕捉人类情感表达的复杂性和多维特征 [5], [6], [7], [8]。例如，在日常对话中，情绪不仅通过语言内容传达，也通过面部表情和声音语调传达；这些模态之间的互补性有助于模型更好地理解情感动态 [9], [10], [11]。

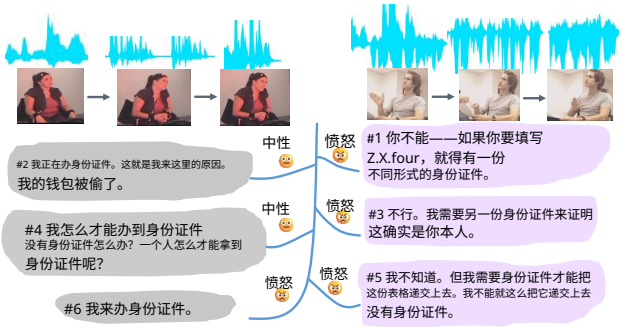


图 1：展示 IEMOCAP 数据集（Ses01F_impro01）中对话动态演变的真实且具有代表性的片段。

值得进一步关注的是，对话中的情感状态会随着互动过程持续演变，呈现出显著的动态特征 [12], [13], [14]。在真实场景中，个体可能维持先前的情感状态，也可能因他人的影响而即时调整，这体现了说话者内连续性和说话者间交互性 [15], [16], [17]。与此同时，不同模态的重要性会在对话过程中动态变化，进而影响模型捕捉情感变化的能力 [18], [19], [20]。图 1 展示了一段典型的对话。男性作为一名

*我们的代码已公开发布于 https://github.com/wys-ljq/AMB-DSGDN。*电子邮箱：20234261@csuft.edu.cn, shouyuntao@stu.xjtu.edu.cn, tanyilong@csuft.edu.cn, aiwei@hnu.edu.cn, mengtao@hnu.edu.cn, lik@newpaltz.edu

*通讯作者 ORCID:

工作人员，从一开始就表现出严厉的态度，其愤怒主要通过语气和面部表情体现，并在随后的交流中持续存在，反映了说话人内依赖的连续性。女性起初以自然的语气提出请求，文本模态提供了理性的语义信息；但随着男性持续表现严厉，她转而变得愤怒，声音强度和面部表情的变化也愈发明显，表明她对他人的情绪十分敏感。这一过程说明，女性的情绪变化更多地受到说话人间情绪影响，凸显了跨说话人依赖关系的动态传染性。与此同时，这段对话也揭示了不同模态在对话过程中的阶段性主导作用：在早期阶段，文本模态通常提供更清晰的语义信息；而当情绪变化显现时，音频和视觉模态则发挥更大的作用。

不同于传统的静态特征建模，多模态对话情感识别任务需要捕捉情绪依赖关系随时间的动态演化 [21], [22], [23]。如果仅采用静态图结构进行建模，情绪依赖关系通常会被简化为固定的关系模式，难以捕捉由上下文变化驱动的情绪动态变化，从而限制对复杂对话场景的建模能力[24], [25], [26], [27]。与此同时，多模态特征通常包含冗余噪声和共享噪声，若将其不加区分地纳入建模过程，可能会掩盖有效的情绪信号，降低表征的判别能力 [28], [29], [30]。此外，该任务还必须应对不同模态在整个对话过程中的贡献波动。如果模型忽视模态之间的动态差异，就可能导致主导模态对结果产生过度影响，而较弱模态无法有效发挥作用，最终降低整体识别性能 [31], [32], [33]。因此，为了有效捕捉对话中持续演化的情感状态，必须实现情绪依赖关系的动态建模、对建模特征中噪声的有效抑制，以及对模态贡献的自适应调节。

为此，多模态对话情感识别（MERC）领域的研究者提出了多种建模方法，主要分为两个方向：序列结构方法依靠循环网络或记忆网络来描述时间依赖关系，能够捕捉局部连续性，但难以覆盖远距离的说话人间交互 [34], [32], [35]；图结构方法构建对话图以建模说话人内部及说话人间的依赖关系，但大多使用静态边权重，忽略了依赖关系随时间和上下文动态变化的特点 [36]。与此同时，面对对话过程中不同模态贡献的波动，现有方法缺乏有效的平衡调节机制，容易导致主导模态占据优势，而较弱的模态则被削弱 [37]。

为解决上述局限，本文提出一种具有自适应模态平衡的动态语义图差分网络。具体而言，为建模情绪依赖关系，我们针对文本、语音和视觉模态构建模态特定子图，以捕捉说话人间交互和说话人内部的情绪演变。在这些子图的基础上，我们设计了一种图差分注意力机制。该机制首先将话语特征投影到统一的表征空间，并通过左右线性变换计算节点之间的注意力分布，同时结合关系嵌入，明确建模说话人间交互。随后，对这两个注意力分布执行差分运算，抵消其重叠部分，仅保留差异部分。该方法能够有效去除不同模态间的共享噪声，同时突出模态特定且与上下文相关的依赖信号。通过这种差分建模策略，模型不仅能过滤冗余信息，还能捕捉对话中随上下文演变的动态依赖关系，从而获得更纯净且更具区分性的情感表示。针对模态调节，我们引入一种自适应模态丢弃机制。模型根据各模态在情感识别中的相对表现计算丢弃概率，并通过概率采样随机丢弃部分主导模态特征，同时对保留的特征进行缩放，以确保整体信息量的稳定性。该策略能够有效缓解模态不平衡，防止任何单一模态在融合过程中占据过强主导地位。通过这两种机制的协同作用，模型能够动态捕捉对话中的情绪演变。本文的主要贡献如下：

- 我们提出了 AMB-DSGDN，通过显式构建模态特定子图，同时建模说话者内部和说话者之间的情感依赖关系。通过联合整合差分图注意力和自适应模态平衡机制，该模型能够有效捕捉对话中的动态情感变化，同时缓解噪声干扰与模态不平衡问题，从而增强情感表示的判别性。

- 我们设计了一种差分图注意力机制，用于计算模态特定子图上成对注意力图之间的差异。通过差分对比，该机制能够抑制共享噪声，突出模态特定信息和与上下文相关的信息，从而提升动态情感建模能力。

- 我们进一步引入一种基于自适应丢弃的模态平衡机制，该机制会动态识别主导模态并随机丢弃其部分特征，同时按比例重新缩放

保留的特征，从而减轻单一模态占主导的影响，并实现平衡的多模态信息融合。

- 在IEMOCAP和MELD数据集上的大量实验证明了AMB-DSGDN的卓越性能和鲁棒性。

2. 相关工作

本节综述多模态学习、动态情感依赖关系以及模态不平衡学习的最新进展，并指出改进多模态对话情绪识别的关键挑战与研究方向。

2.1. 多模态学习

近年来，多模态学习在计算机视觉、自然语言处理和语音分析等领域取得了显著进展。研究人员提出了多种方法，以增强多源信息融合并提升模型的泛化性能。在多模态情感识别任务中，NORM-TR模型 [38] 有效捕捉各模态之间的长程依赖关系，并通过抗噪声特征提取器与噪声感知学习方案结合Transformer融合机制，提升了模型对噪声的鲁棒性和计算效率。Affect-GPT模型 [39] 采用预融合操作和多模态大语言模型架构，将跨模态交互置于大语言模型之外，从而从文本、音频和视觉输入中捕捉细粒度情绪。HKD-MER模型 [40] 通过层次化知识蒸馏将主导模态知识迁移至其他模态，从而增强特征平衡性与判别能力。MERBench模型 [41] 利用基于注意力的融合框架，实现对模态异质性与跨模态交互的细粒度建模。总体而言，这些方法在融合效果和泛化能力方面取得了积极成果。然而，现有研究往往未能充分考虑对话场景中各模态情感表达的动态差异，无法有效滤除多模态特征中的冗余噪声，并且主导模态（如文本）过度主导融合过程，抑制了非主导模态的互补作用，从而限制了整体识别性能。为此，本文构建用于动态情感建模的模态特定子图，引入差分图注意力以抵消共享噪声，并采用自适应 dropout 策略调整模态贡献，从而改进多模态模型中的联合学习。

2.2. 动态情感依赖关系

在情感识别领域，近期研究提出了多种方法，以更好地捕捉动态情感依赖关系，主要包括两类：基于循环神经网络（RNN）和基于图卷积网络（GCN）的模型。基于 RNN 的方法（例如优化的 RNN 模型 [42]）能够处理序列数据并建模时间依赖关系，但在长序列中容易出现梯度消失或梯度爆炸，从而限制了对复杂情感动态的表达能力。为增强建模能力，研究人员使用 GCN 处理情感关联结构，适用于多轮交互的对话场景。DER-GCN [43]通过融合对话感知和事件感知信息，强化了说话人间依赖关系的建模；SERC-GCN[44]捕捉说话人的情感状态变化，从而更好地刻画情感演变。一些研究也尝试融合 RNN 和 GCN，例如 GCN-LSTM [45],, 以平衡时间信息和结构信息。此外，DEDNet[46]对说话人间和说话者内的情感依赖关系进行建模，并利用交互捕捉情感变化。尽管这些方法取得了进展，但大多数基于图的方法仍依赖静态边权重，而且多模态特征中的噪声会削弱对动态情感依赖关系的表征。因此，本文在每种模态中构建说话者内子图和说话人间子图，以表达情绪的时间关系和交互关系。同时，在图中引入正负分支差分注意力，利用两者注意力的差异并增强稳定且一致的情感特征，从而实现更可靠的动态情感建模。

2.3. 模态不平衡学习

近年来，模态不平衡是多模态学习面临的核心挑战之一。不同模态在数据质量、信息密度和可用性方面存在差异，导致某些模态在融合过程中占据主导地位，而其他模态的贡献则被削弱。为解决这一问题，研究人员提出了多种策略。Wei等人 [47]设计了一种动态模态分配框架，能够自适应地调整每种模态的权重，增强次要模态的贡献，缓解单一模态主导的问题。Wang等人 [48] 提出了一种对抗式模态平衡方法，通过数量—质量重加权增强表现不足的模态特征，改善融合效果。Guo等人提出的MPLMM模型 [49] 在情感识别过程中动态抑制噪声模态，同时保留关键信息。Lian等人提出的MER框架 [50] 将模态鲁棒性与半监督学习相结合，增强次要模态的判别能力。Chen等人 [51]在教师—学生架构中采用多模态知识蒸馏，实现跨模态知识迁移，缓解模态不平衡导致的性能下降。然而，现有方法未能充分考虑动态对话中模态贡献随上下文变化而产生的波动，导致模型在极端不平衡情况下出现更大的波动。为此，本文提出一种自适应模态平衡策略，通过评估每种模态在当前批次中的表现，动态调整特征采样，并按比例放大保留的特征，从而在主导模态过于

强势时适度抑制其特征，同时增强次要模态的贡献，实现多模态特征的平衡融合。

3. 任务定义

对话多模态情绪识别旨在识别对话序列中每个话语的情绪类别。该任务的输入是由多个话语组成的对话序列，其中每个话语都包含文本、视觉和音频模态的特征，以及说话人信息和上下文关系。形式化地，给定一个对话序列 $\mathcal{D} = \{u_1, u_2, \ldots, u_N\}$，其中 N 是话语数量，每个话语 u_i 由其文本特征 $\mathbf{t}_i \in \mathbb{R}^{d_t}$、视觉特征 $\mathbf{v}_i \in \mathbb{R}^{d_v}$、音频特征 $\mathbf{a}_i \in \mathbb{R}^{d_a}$、说话人标识符 s_i 和情绪标签 $y_i \in \{1, 2, \ldots, C\}$ 组成（C 是情绪类别的数量）。模型的目标是学习映射函数 $f : \mathcal{D} \to \{y_1, y_2, \ldots, y_N\}$，以最大限度地提高预测准确率和加权F1分数。该任务融合了多模态融合、说话人依赖关系和上下文关系，因此适用于 IEMO-CAP 和 MELD 等数据集。

4. 方法

本节介绍所提出的对话多模态情感识别方法。该方法包含六个关键组成部分：模型架构、话语级编码器、关系子图构建、差分注意力图卷积网络、动态模态平衡机制和情感分类器。

4.1. 模型架构

AMB-DSGDN 将差分注意力图卷积网络与动态模态平衡机制相结合。图 2 展示了该模型的整体框架，主要由五个核心组件构成：话语级编码器、关系子图构建、差分注意力图卷积网络、动态模态平衡机制和情感分类器。首先，文本、音频和视觉特征通过各自的线性变换被映射为统一维度的隐藏表示。随后，引入融合说话人嵌入的 Transformer 编码器进行上下文建模，以捕捉话语之间的时间依赖关系。在此基础上，该模型构建说话人内关系子图和说话人间关系子图，并利用差分注意力图卷积机制对情感依赖进行建模。该机制通过建模同一模态子图中节点注意力分布的差异，抑制共享冗余与噪声模式，从而突出真实的情感依赖信号。这一设计有助于更准确地刻画对话中的动态情感变化。在模态调节方面，自适应模态丢弃机制根据各模态的贡献生成丢弃概率于情感建模，并对补偿后的保留特征进行缩放，以实现跨模态的动态权重调整。最后，将融合后的多模态特征输入情感分类器进行预测，同时引入辅助损失，以增强单模态情绪表征的鲁棒性。

4.2. 话语级编码器

话语级编码器用于对多模态输入（文本、视觉和音频）进行特征提取与融合，为对话中的每个话语生成语义表示。该模块由模态特征投影、位置编码、说话人嵌入和 Transformer 编码层组成，最终输出三种模态的统一表征。

对于对话中的每个话语，首先从预训练模型中提取初始特征：文本模态使用 RoBERTa [52]，视觉模态使用 DenseNet [53]，音频模态使用 OpenSmile [54]。随后，通过线性映射将不同模态的特征投影到统一的隐藏维度：

$$\mathbf{x}_i^t = W_t \cdot \mathbf{f}_i^t + \mathbf{b}_t, \quad \mathbf{x}_i^v = W_v \cdot \mathbf{f}_i^v + \mathbf{b}_v, \quad \mathbf{x}_i^a = W_a \cdot \mathbf{f}_i^a + \mathbf{b}_a, \tag{1}$$

其中，$\mathbf{f}_i^t$、$\mathbf{f}_i^v$、$\mathbf{f}_i^a$ 分别表示第 i 个话语的原始文本、视觉和音频特征；$W_t \in \mathbb{R}^{d_h \times D_t}$、$W_v \in \mathbb{R}^{d_h \times D_v}$、$W_a \in \mathbb{R}^{d_h \times D_a}$ 为投影矩阵；$\mathbf{b}_t$、$\mathbf{b}_v$、$\mathbf{b}_a$ 为偏置项。

为引入序列位置信息和说话人身份信息，模型添加了位置编码和说话人嵌入。位置编码采用正弦余弦形式：

$$\begin{aligned} PE(pos, 2j) &= \sin\left(\frac{pos}{10000^{2j/d_h}}\right), \\ PE(pos, 2j+1) &= \cos\left(\frac{pos}{10000^{2j/d_h}}\right) \end{aligned} \tag{2}$$

其中，pos 表示该话语在序列中的位置，j 为维度索引。说话人嵌入通过嵌入层生成：$\mathbf{s}_i = E_s(\text{spk}_i)$，其中 $E_s \in \mathbb{R}^{(n_s+1) \times d_h}$、$n_s$ 为说话人数。

对于音频和视觉模态，将投影后的特征与位置编码和说话人嵌入相加：

$$\mathbf{x}_i^a = \mathbf{x}_i^a + PE(i) + \mathbf{s}_i, \quad \mathbf{x}_i^v = \mathbf{x}_i^v + PE(i) + \mathbf{s}_i. \tag{3}$$

随后，将音频和视觉模态分别输入独立的 Transformer编码器，以建模上下文依赖关系。每个编码器由一个单层多头自注意力层和一个逐位置前馈网络组成。自注意力机制定义如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{4}$$

其中，$Q = W_q\mathbf{x}$、$K = W_k\mathbf{x}$、$V = W_v\mathbf{x}$、$d_k = d_h/h$、h 为注意力头的数量。逐位置前馈网络定义为：

$$FFN(\mathbf{x}) = W_2 \cdot \text{GELU}(W_1 \cdot LN(\mathbf{x})) + \mathbf{x}, \tag{5}$$

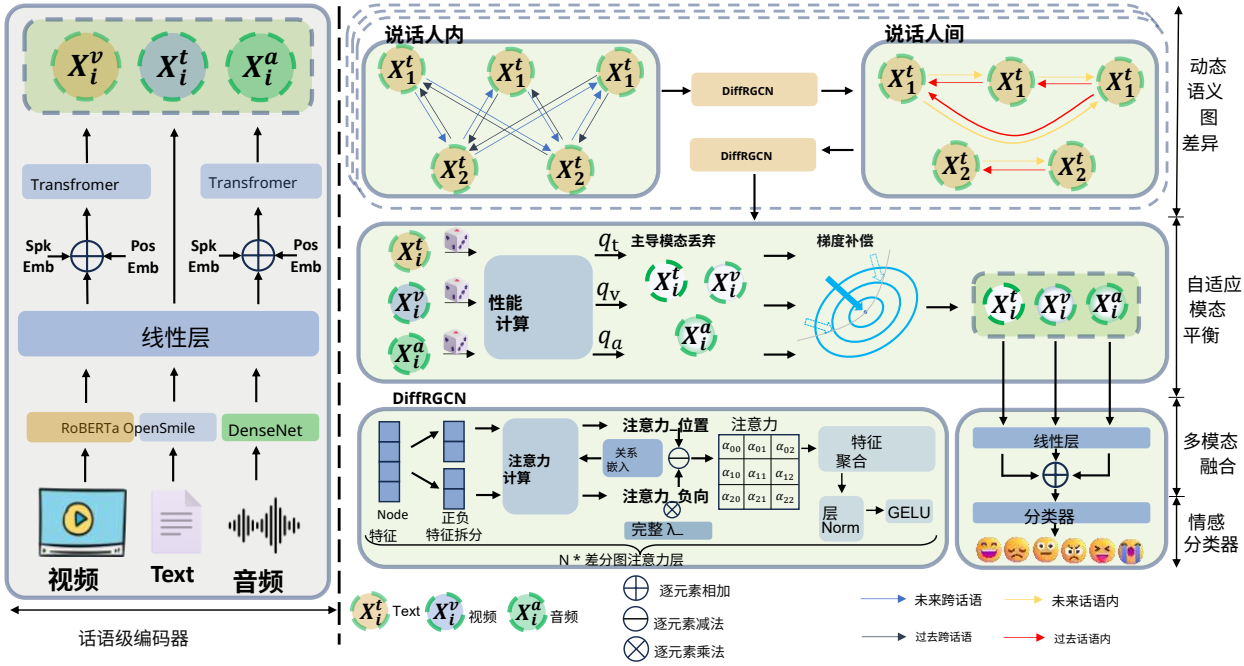


图 2： 该架构包含四个核心模块：首先，话语级编码器通过 OpenSmile（音频）、RoBERTa（文本）和 DenseNet（视频）提取单模态特征，并在 Transformer 编码中结合说话人嵌入（Spk Emb）和位置嵌入（Pos Emb），获得第 i个话语的文本特征$\mathbf{x}_i^t$、视频特征$\mathbf{x}_i^t$和音频特征$\mathbf{x}_i^t$；其次，差分图注意力模块为每种模态构建子图，包括“说话人内子图（Intraspeaker）”和“说话人间子图（Interspeaker）”，通过差分图注意力计算两组注意力分布之间的差异，消除跨模态共享噪声并保留模态特定情感信号；然后，自适应模态平衡模块根据批次级性能计算每种模态的丢弃概率（$q_t/q_v/q_a$），对占主导地位的模态执行动态丢弃，同时通过梯度补偿缩放保留的特征，以维持信息平衡；最后，多模态融合与分类模块通过线性层融合平衡后的特征，并将其输入分类器，获得情感识别结果。

其中， LN 表示层归一化。

最后，音频和视觉模态的上下文表示统一表示为：

$$\mathbf{x}_i^m, \mathbf{A}_{attn}^m = \text{TransformerEncoder}(\mathbf{x}^m, \mathbf{mask}, \mathbf{s}), \quad m \in \{a, v\}, \tag{6}$$

其中， $\mathbf{A}_{attn}^m$ 是反映话语间依赖关系的注意力矩阵， $\mathbf{mask}$ 是话语掩码， $\mathbf{s}$ 表示序列信息。

4.3. 关系子图构建

在对话中，每个话语 u_i 都表示为图节点 v_i，其多模态特征为$\mathbf{x}_i^m$， $m \in \{t, v, a\}$，分别对应文本、视觉和音频模态。

4.4. 图构建与表示

完整对话表示为$U = \{u_1, u_2, \ldots, u_{N_b}\}$，并建模为有向图 $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$，其中每个话语 u_i 对应于节点 $v_i \in \mathcal{V}$。节点特征由多模态表示 $(\mathbf{x}_i^t, \mathbf{x}_i^v, \mathbf{x}_i^a)$组成，分别对应文本、视觉和音频模态。为建模情感依赖关系，该模型构建两个关系子图：说话人内子图 $\mathbf{Adj}_s$ 和说话人间子图 $\mathbf{Adj}_c$，分别刻画时间演化，其范围在同一说话人内部以及不同说话人之间的交互关系。边集 $\mathcal{E}$ 表示话语之间的依赖关系，边类型 $\mathcal{R}$ 将时间与情感关系编码为五类：说话人自身过去、说话人自身未来、说话人间过去、说话人间未来和自环。边权重 $\mathcal{W}$ 表示交互强度或邻近度。邻接范围受窗口大小 $w = 5$ 约束，以减少长距离噪声和计算复杂度。在批处理过程中，对话最大长度设为 $L = \max_b N_b$，填充位置置为零。两个子图被堆叠为 $\mathbf{Adj}_s$ 和 $\mathbf{Adj}_c$，为后续图注意力建模提供结构信息。

说话人内子图 $\mathbf{Adj}_s$ 仅连接来自同一说话人的话语，用于建模内部一致性和时间演变：

$$\}, \quad (\mathbf{Adj}_s)_{i,j} = \begin{cases} 1 & \text{if } i = j \wedge s_i = s_j, \\ 2 & \text{if } i > j \wedge |i-j| \le w \wedge s_i = s_j \quad \text{(intra-past)}, \\ 3 & \text{if } i < j \wedge |i-j| \le w \wedge s_i = s_j \quad \text{(intra-future)}, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

这一设计突显了过去话语对当前情绪的累积影响，同时引入了来自未来话语的前瞻性上下文。

说话人间子图 $\mathbf{Adj}_c$ 用于对不同说话人之间的动态交互关系进行建模：

$$(\mathbf{Adj}_c)_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 4 & \text{if } i < j \wedge |i-j| \le w \wedge s_i \neq s_j \quad \text{(未来跨说话人)}, \\ 5 & \text{if } i > j \wedge |i-j| \le w \wedge s_i \neq s_j \quad \text{(过去跨说话人)}, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

该子图刻画说话人间情感动态中的回应、冲突与协作。

最终将两个关系子图表示为$\mathbf{Adj}_s, \mathbf{Adj}_c \in \mathbb{Z}^{B\times L\times L}$，其中 $r_{i,j} \in \mathcal{R}$，而 $\mathbf{Adj}_{i,j} = \mathrm{id}(r_{i,j})$。分离的子图结构支持分层建模：首先聚合说话人之间的全局交互信息，然后细化单个说话人的内部表示，从而增强模型在多说话人场景中的建模能力。

上述关系子图在不同模态之间共享，但作用于各自的模态表征。具体而言，文本、视觉和音频模态使用独立的差分图注意力网络，并在相同的 $\mathbf{Adj}_s$ 和 $\mathbf{Adj}_c$ 上进行建模。以文本模态为例，说话人间图注意力网络（‘gatTer‘）首先作用于 $\mathbf{Adj}_c$，随后说话者内图注意力网络（‘gatT‘）作用于 $\mathbf{Adj}_s$。音频和视觉模态分别使用‘gatAer‘ / ‘gatA‘和‘gatVer‘ / ‘gatV‘。该设计使各模态能够在共享对话结构的同时独立学习关系表示。

4.5. 差分注意力图卷积网络

为融合关系子图中的结构信息，我们提出差分关系图卷积网络（DiffRGCN）。该网络引入基于图注意力网络的差分注意力机制，以建模关系类型和特征差异。DiffRGCN由多头注意力层和输出层组成，其中每个注意力头通过差分机制隐式刻画节点之间的增强与抑制关系，并结合关系标签增强语义感知。该网络可独立应用于不同模态，并依次作用于说话人间子图和说话人内子图，先聚合说话人间交互，再细化说话人内表征，具体过程如算法1。

DiffRGCN 的核心是差异图注意力层。该层将输入特征分为正分支和负分支，分别建模情绪增强和抑制信号，并实现差异

算法1： 差分图注意力层

输入： 节点特征 $\mathbf{H}^m$，邻接矩阵 $\mathbf{Adj}$，层深度 d，关系感知标志

使用公式（9）投影节点特征，得到 $\mathbf{Wh}$；
将 $\mathbf{Wh}$ 拆分为正分支和负分支；
使用以下内容生成 $\mathbf{q}^{\text{pos}}$，$\mathbf{k}^{\text{pos}}$，$\mathbf{v}^{\text{pos}}$ 和 $\mathbf{q}^{\text{neg}}$，$\mathbf{k}^{\text{neg}}$，$\mathbf{v}^{\text{neg}}$ 公式 (10) 和公式 (11)；
如果 启用关系感知**则**
    将邻接关系嵌入注意力分数中；
使用 $\mathbf{e}^{\text{pos}}$ 和 $\mathbf{e}^{\text{neg}}$ 计算原始注意力分数 公式 (12)；
应用激活函数、掩码和 Softmax 归一化以获得 $\boldsymbol{\alpha}^{\text{pos}}$ 和 $\boldsymbol{\alpha}^{\text{neg}}$；
计算差分系数 λ_1、 λ_2和 λ_{full} 使用公式(13)和公式(14)；
融合正负注意力权重以 使用公式(15)获得最终的 $\boldsymbol{\alpha}$；
拼接值向量 $\mathbf{v}^{\text{pos}}$ 以及 $\mathbf{v}^{\text{neg}}$ 以获得 $\mathbf{Wh}_v$（公式（16））；
聚合 $\mathbf{Wh}_v$，使用注意力权重 $\boldsymbol{\alpha}$ 以获得 $\mathbf{h}'$（公式（17））；
拼接多头输出，得到最终节点表征 $\mathbf{x}$（公式（18））；
返回 $\mathbf{x}$ 和注意力权重 $\boldsymbol{\alpha}$；

通过可学习的 lambda 参数进行聚合。具体而言，输入特征首先经过线性投影：

$$\mathbf{Wh} = \mathbf{H}^m\mathbf{W}, \tag{9}$$

其中，$\mathbf{H}^m \in \mathbb{R}^{B\times L\times d_{\text{in}}}$、$\mathbf{W} \in \mathbb{R}^{d_{\text{in}}\times d_{\text{out}}}$、$B$ 表示批次大小，L表示节点数。

在正分支中，投影特征用于生成查询向量和键向量：

$$\mathbf{q}^{\text{pos}} = \mathbf{a}^{\text{pos}}_{\text{left}}\left(\mathbf{Wh}^{\text{pos}}\right), \quad \mathbf{k}^{\text{pos}} = \mathbf{a}^{\text{pos}}_{\text{right}}\left(\mathbf{Wh}^{\text{pos}}\right). \tag{10}$$

负分支相应地生成 $\mathbf{q}^{\text{neg}}$ 和$\mathbf{k}^{\text{neg}}$。

正分支和负分支的值向量定义如下：

$$\mathbf{v}^{\text{pos}} = \mathbf{Wh}^{\text{pos}}, \quad \mathbf{v}^{\text{neg}} = \mathbf{Wh}^{\text{neg}}. \tag{11}$$

启用关系感知后，邻接矩阵被嵌入为关系表示 $\mathbf{R} \in \mathbb{R}^{B\times L\times L\times d_r}$，并投影得到 $\mathbf{rel}^{\text{pos}}$ 和 $\mathbf{rel}^{\text{neg}}$，再将其加到基础注意力分数上：

$$\begin{aligned} \mathbf{e}^{\text{pos}}_{i,j} &= \mathbf{q}^{\text{pos}}_i + \left(\mathbf{k}^{\text{pos}}_j\right)^\top + \mathbf{rel}^{\text{pos}}_{i,j}, \\ \mathbf{e}^{\text{neg}}_{i,j} &= \mathbf{q}^{\text{neg}}_i + \left(\mathbf{k}^{\text{neg}}_j\right)^\top + \mathbf{rel}^{\text{neg}}_{i,j}. \end{aligned} \tag{12}$$

正向注意力分数经 LeakyReLU 激活后，使用 $\mathbf{Adj} \le 0$ 屏蔽无效边，再通过 Softmax 归一化，得到正向注意力权重$\boldsymbol{\alpha}^{\text{pos}}$。负向分支采用相同的处理流程，得到$\boldsymbol{\alpha}^{\text{neg}}$。

为平衡正向与负向注意力，引入差异 lambda 参数。首先，我们计算：

$$\lambda_1 = \exp\left(\sum_{d=1}^{D} \lambda_{\text{left},1}^{(d)} \cdot \lambda_{\text{right},1}^{(d)}\right), \\ \lambda_2 = \exp\left(\sum_{d=1}^{D} \lambda_{\text{left},2}^{(d)} \cdot \lambda_{\text{right},2}^{(d)}\right), \tag{13}$$

其中，所有参数均初始化自均值为0、方差为0.1的正态分布。最终的差分系数定义为：

$$\lambda_{\text{full}} = \lambda_1 - \lambda_2 + \lambda_{\text{init}}, \tag{14}$$

其中，$\lambda_{\text{init}} = 0.8 - 0.6\exp(-0.3 \cdot \text{深度})$，depth表示层深度。

最终的注意力权重通过差异融合获得：

$$\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\text{pos}} - \lambda_{\text{full}} \cdot \boldsymbol{\alpha}^{\text{neg}}. \tag{15}$$

该操作增强有效的情感依赖，同时抑制噪声干扰。

随后，将正分支和负分支的值向量进行拼接：

$$\mathbf{Wh}_v = \text{cat}(\mathbf{v}^{\text{pos}}, \mathbf{v}^{\text{neg}}) \in \mathbb{R}^{B\times L\times d_{\text{out}}}, \tag{16}$$

并在注意力权重下完成邻居特征聚合：

$$\mathbf{h}' = \boldsymbol{\alpha}\mathbf{Wh}_v. \tag{17}$$

DiffRGCN 采用多头差分注意力并行建模多视图关系，并将输出拼接为：

$$\mathbf{x} = \text{cat}\big([\text{att}_i(\mathbf{x}, \mathbf{Adj})]\big). \tag{18}$$

拼接结果依次经过丢弃、输出注意力层和带残差连接的全连接层，最后通过层归一化获得节点表示。

该设计将多头机制与差异建模相结合，在保持计算复杂度 $O(BLwd \cdot h + BL^2/h)$可控的同时，有效增强了关系建模的表达能力和稳定性。

4.6. 自适应模态平衡

在多模态学习中，各模态对情感建模的贡献通常不平衡，这可能导致弱模态在融合过程中被忽视。为解决该问题，我们提出一种动态模态平衡机制，通过自适应模态丢弃策略动态调整模态权重。详细流程如算法2所示。为确保及时响应模态贡献的变化，模态性能指标 p_m、比率参数 $r_{m,j}$ 以及

算法 2： 自适应模态丢弃

输入： 模态特征列表 $\mathbf{F} = [\mathbf{T}, \mathbf{V}, \mathbf{A}]$，批次大小 B，基础丢弃概率 q_{base}，缩放因子 λ，标签 $\mathbf{y}$，训练掩码 $\mathbf{U}$

对于 每个训练批次**执行**
- // 计算模态性能 p_m
- **对于** 每个模态 m **执行**
  - $p_m = \text{F1}(\mathbf{F}_m, \mathbf{y}, \mathbf{U})$;
- // 计算相对性能比率和丢弃概率
- **对于**每种模态 m **执行**
  - $r_{m,j} = \frac{p_m}{p_j+\epsilon} - 1, \forall j \neq m$;
  - $q_m = \text{裁剪}(q_{\text{base}} \cdot (1 + \lambda \cdot \text{Softmax}(\text{ReLU函数}(r_{m,j}))), 0, 1)$;
- // 应用丢弃和缩放
- 随机决定是否应用丢弃;
- **如果**应用了丢弃**则**
  - 采样伯努利掩码 $\mathbf{M} \sim 1 - q$;
  - 应用掩码：$\mathbf{F}' = \mathbf{M} \odot \mathbf{F}$;
  - 按期望缩放：$\mathbf{F}'' = \mathbf{F}'/(1-\theta)$;
  - 保留至少包含一个模态的实例：$\mathbf{F}''' = \mathbf{F}''[\mathbf{u}]$;
- **else**
  - $\mathbf{F}''' = \mathbf{F}$，$\mathbf{u} = \mathbf{1}$;

返回 丢弃并缩放的模态特征 $\mathbf{F}'''$和有效实例掩码 $\mathbf{u}$。

丢弃概率 q_m 在每个训练批次中更新。该机制仅在模型进入稳定训练阶段后激活。稳定训练阶段的判定依据预热轮数的敏感性实验结果。如第 6.2 节中的实验所示，模型在 IEMOCAP 数据集上经过约 60 个训练轮次后性能趋于稳定并达到最优水平。因此，本文将预热轮数设为约 60，以避免早期参数波动对模型收敛的影响。

首先，我们计算每个实例的模态级性能指标 p_m。我们使用加权F1分数^进行测量：$p_m = \text{F1}(\hat{y}_m, y, \mathbf{U})$，其中 $\hat{y}_m = \arg\max(\mathbf{P})$, y表示真实标签，而 $\mathbf{U}$ 是用于筛除填充位置的话语掩码。其定义如下：

$$p_m = \frac{\sum_{c=1}^{C} w_c \cdot \text{Precision}_c \cdot \text{Recall}_c}{\sum_{c=1}^{C} w_c}, \tag{19}$$

其中，$w_c = \frac{N}{\sum_i \mathbb{I}(y_i=c)}$ 是类别 c 的逆频率权重。精确率和召回率根据有效话语计算。具体而言，我们首先展平 logit 和标签：$\mathbf{P}_m^{\flat} = \mathbf{P}_m[\mathbf{U} == 1]$、$y^{\flat} = y[\mathbf{U} == 1]$，然后计算精确率$_c$ = TP$_c$ TP$_c$+FP$_c$、召回率$_c$ = TP$_c$ TP$_c$+FN$_c$。

随后，我们计算不同模态之间的相对性能差异。对于模态 m，其与其他模态的性能比定义为：

$$r_{m,j} = \frac{p_m}{p_j + \epsilon} - 1, \quad \forall j \neq m, \tag{20}$$

其中，$\epsilon = 10^{-5}$ 是用于避免除以零的平滑因子。此外，对于任意模态 m，其相对性能比率向量 $\mathbf{r}_m$ 的维度为 $|\mathcal{M}| - 1$，其中$\mathcal{M} = \{t, v, a\}$。$\mathbf{r}_m$的每个维度对应于模态 m 与另一模态 $j(j \neq m)$之间的性能比率 $r_{m,j}$，即：

$$\mathbf{r}_m = [r_{m,j}]_{j \in \mathcal{M}, j \neq m}. \tag{21}$$

在后续计算中，我们不显式筛选 $\mathbf{r}_m$中的正值，而是直接对完整的比率向量应用 ReLU 激活，以抑制负项，并对非负结果执行 Softmax 归一化，从而平滑模态之间的相对性能差异。具体而言，我们按索引顺序排列所有相对性能比率 $r_{m,j}(j \in \mathcal{M}, j \neq m)$，构建比率向量 $\mathbf{r}_m$，首先施加非负约束：$\mathbf{r}_m^+ = \max(\mathbf{r}_m, 0)$，然后执行 Softmax 归一化：

$$\hat{\mathbf{r}}_m = \frac{\exp(\mathbf{r}_m^+)}{\sum \exp(\mathbf{r}_m^+)}, \tag{22}$$

最终得到加权平均比率：

$$\bar{r}_m = \frac{\sum \hat{\mathbf{r}}_m \odot \mathbf{r}_m}{|\mathbf{r}_m|}. \tag{23}$$

基于该比率，模态丢弃概率定义为：

$$q_m = q_{\text{base}} \cdot (1 + \lambda \cdot \bar{r}_m), \quad q_m = \text{clip}(q_m, 0, 1), \tag{24}$$

其中 $q_{\text{base}} = 0.3$ 和 $\lambda = 0.9$ 是超参数。该设计使性能较高的模态具有更高的丢弃概率，从而鼓励模型关注较弱的模态，同时避免极端差异主导训练过程。

执行模态丢弃时，对于模态特征集 $\mathbf{F} = [\mathbf{T}, \mathbf{V}, \mathbf{A}] \in \mathbb{R}^{M \times B \times N \times d_h}$（其中 $M = 3$ 和 B 是实例数），我们首先以 $p_{\text{exe}} = 0.5$ 的概率决定是否执行丢弃。如果执行丢弃，则生成一个伯努利掩码：

$$\mathbf{M}_{m,b} \sim \text{Bernoulli}(1 - q_m), \tag{25}$$

并将其应用于特征：$\mathbf{F}' = \mathbf{M} \odot \mathbf{F}$。

随后，我们通过自定义的 Autograd 函数进行期望补偿缩放：

$$\mathbf{F}'' = \frac{\mathbf{F}'}{1 - \theta}, \quad \theta = \frac{\sum_m d_m q_m}{\sum_m d_m}, \tag{26}$$

其中，$d_m = d_h$ 为模态维度。其前向传播采用线性缩放，反向传播则仅将梯度传递至输入：

$$\frac{\partial \mathbf{F}''}{\partial \mathbf{F}'} = \frac{1}{1 - \theta}, \quad \frac{\partial \mathbf{F}''}{\partial \theta} = 0. \tag{27}$$

最后，我们使用更新标志$\mathbf{u}_b = \sum_m \mathbf{M}_{m,b} > 0$筛选有效实例，仅保留至少具有一个模态的样本：$\mathbf{F}''' = \mathbf{F}''[\mathbf{u}]$，从而避免因完全丢弃模态而导致的空表示问题。

4.7. 情感分类器

获得多模态动态表示后，我们设计了一个多模态融合与分类模块，用于整合文本、视觉和音频特征并生成情绪预测。对于每种模态的最终表示，模型使用独立的分类头进行投影。每个分类头由 ReLU 激活、Dropout 层和线性映射组成，用于将隐藏表示映射到类别空间。以文本模态为例，其分类过程定义如下：

$$\mathbf{t} = \sigma(\text{Dropout}(\text{ReLU}(\mathbf{x}_i^t \mathbf{W}_t + \mathbf{b}_t))). \tag{28}$$

其中，$\mathbf{W}_t \in \mathbb{R}^{d_h \times C}$、$\mathbf{b}_t \in \mathbb{R}^C$、$\sigma$表示线性映射后的输出。视觉和音频模态的分类头 $\mathbf{v}$ 和 $\mathbf{a}$ 以相同方式构建。

随后，对每种模态的 logit 进行逐元素相加，以完成融合：

$$\mathbf{O} = \mathbf{t} + \mathbf{v} + \mathbf{a}, \tag{29}$$

并使用对数 softmax 计算融合概率分布：

$$\mathbf{P} = \log(\text{softmax}(\mathbf{O})), \quad \hat{\mathbf{P}} = \text{softmax}(\mathbf{O}). \tag{30}$$

$\hat{}$通过对 $\mathbf{P}$取最大值索引获得预测标签。为增强单模态表示能力，模型同时计算单模态对数 softmax 概率$\mathbf{P}_t$、$\mathbf{P}_v$ 和 $\mathbf{P}_a$，用于辅助优化。

模型训练采用交叉熵损失，并结合序列掩码来适应变长输入。融合损失定义如下：

$$\mathcal{L}_{\text{fusion}} = \text{CE}(\mathbf{P}, \mathbf{y}, \mathbf{m}), \tag{31}$$

其中，交叉熵形式如下：

$$\text{CE}(\mathbf{P}, \mathbf{y}, \mathbf{m}) = -\sum_{i=1}^{L} m_i \sum_{c=1}^{C} y_{i,c} \log P_{i,c}. \tag{32}$$

其中，L 表示序列长度，C 表示类别数，$P_{i,c}$表示预测概率，$y_{i,c}$表示独热标签，$m_i \in 0, 1$表示掩码。最终训练目标由融合损失和单模态辅助损失构成：

$$\mathcal{L} = \mathcal{L}_{\text{fusion}} + \alpha_t \mathcal{L}_t + \alpha_a \mathcal{L}_a + \alpha_v \mathcal{L}_v, \tag{33}$$

其中 $\alpha_t = \mathcal{L}_t / 10$ 等是用于平衡不同模态训练贡献的自适应权重。

表1IEMOCAP 和 MELD 数据集的统计信息，包括对话数、话语数、说话人数、情绪类别总数，以及各情绪类别的话语数。数值按训练集+验证集/测试集划分报告。

数据集	对话数	话语数	说话人	类别数	中性	快乐 Joy	悲伤	愤怒	兴奋 惊讶	沮丧	厌恶	Fear
IEMOCAP	120 / 31	5810 / 1623	2	6	1708 / 490	1636 / 456	1084 / 266	1103 / 290	1041 / 265	1848 / 324	-	-
MELD	1153/ 280	11098/2610	9	7	5180 / 1256	1940 / 368	794 / 208	1243 / 364	1205 / 431	-	293/ 68	268 / 50

5. 实验

本节介绍本研究的实验设置，包括数据集、基线模型、评估指标和实现细节。通过这些设置，我们对所提出的模型进行了全面评估，并将其与现有最先进的方法进行比较，以验证其有效性。

5.1. 数据集

如表 1所示，本实验采用了两个广泛使用的多模态对话情感识别数据集
:
IEMOCAP 和 MELD。这些数据集包含文本、音频和视觉等多模态特征，适用于评估模型在多模态融合和对话上下文建模方面的性能。

IEMOCAP 数据集 [55]：该数据集由南加州大学收集，共包含 10 个对话轮次，总计约 12 小时的视频录制。每段对话涉及两名说话人，并标注了 6 种情绪类别：中性、快乐、悲伤、愤怒、兴奋和沮丧。该数据集共包含 5531 条话语。我们遵循标准划分方法，将前 8 个轮次用作训练集，最后 2 个轮次用作测试集。该数据集的多模态特性使其成为评估情感识别模型的基准。

MELD 数据集 [2]：该数据集基于电视剧《老友记》中的对话片段，包含约 1400 段对话和 13708 个话语。数据集标注了 7 种情绪类别：中性、惊讶、恐惧、悲伤、喜悦、厌恶和愤怒。我们采用官方划分：约 1000 段对话用于训练集，100 段用于验证集，300 段用于测试集。该数据集的说话人数量较多（最多 9 人），增加了对话上下文建模的复杂性。

5.2. 基线

为验证所提出模型的优越性，我们选取了以下基线模型进行比较。这些模型涵盖了基于循环神经网络、图卷积网络和多模态融合的代表性方法。我们在相同的数据集上复现了这些模型，或使用公开代码进行了实验。

DialogueRNN [34]：一种基于循环神经网络的模型，利用注意力机制捕捉对话中的情感动态。

DialogueGCN [36]：一种基于图卷积网络的模型，将对话建模为图结构，以捕捉话语之间的依赖关系。

MMGCN [56]：一种多模态图注意力网络，通过图注意力机制融合文本、音频和视觉模态，以进行情感识别。

MM-DFN [57]：一种利用动态注意力融合多模态特征的多模态动态融合网络。

COGMEN [58]：一种用于多模态情感识别的上下文化图神经网络，结合了 GNN 和多模态融合。

MultiEMO [59]：一种基于注意力的相关性感知多模态融合框架，着重强调模态间的相关性。

SDT [60]：一种依赖说话人的 Transformer 模型，利用 Transformer 捕捉说话人特定模式。

GraphCFC [61]：一种用于多模态情感分析的基于有向图的跨模态融合网络。

RL-EMO [62]：一种经强化学习增强的情感识别模型，通过 RL 优化情感预测。

DEDNet [46]：一种用于处理多模态对话情绪的双编码器-解码器网络。

DER-GCN [43]: 该模型构建加权多关系图，以捕捉对话中的多样化依赖关系，并通过自监督掩码图自编码器和多信息 Transformer 融合多模态特征。

MERC-PLTAF [5]: 该研究采用细粒度特征提取和跨模态融合策略，对跨对话的多模态情感信息进行联合建模。

5.3. 评估指标

我们采用加权F1分数和加权准确率作为主要评估指标，以全面评估模型性能。所有指标均使用 Scikit-learn 库在测试集上计算。

5.4. 实现细节

模型基于 PyTorch 框架实现。对于文本模态，使用 RoBERTa 提取维度为 1024 的特征。对于视觉模态，使用 DenseNet 提取维度为 342 的特征；对于音频模态，使用 OpenSmile 提取特征，IEMOCAP 数据集的特征维度为 1582，MELD 数据集的特征维度为 300。随后，通过线性层将这些初始特征投影到统一的 512 维隐藏维度，以进行后续处理。为确保训练稳定性和模型性能，批次大小设为 16，学习率设为 0.000068，使用带有 0.00005 权重衰减的 Adam 优化器，并训练 100 轮。模态丢弃机制的

表2该表使用 F1分数评估所有模型在 IEMOCAP（六种情绪类别）数据集上的性能，并以加权准确率（wa-ACC）和加权F1分数（wa-F1）为指标，展示模型在三个数据集上的整体性能，其中最佳结果以粗体标示，次佳结果加下划线。

模型	IEMOCAP													
	快乐		sad		中性		愤怒		兴奋		沮丧		加权准确率	加权 F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
DialogueRNN	25.00	34.95	82.86	84.58	54.43	57.66	61.76	64.42	**90.97**	76.30	62.20	59.55	65.43	64.29
DialogueGCN	64.29	29.03	80.86	64.37	43.14	50.96	68.49	63.29	71.85	68.19	57.68	62.41	62.07	58.19
MMGCN	32.64	39.66	72.65	76.89	65.10	62.81	73.53	71.43	77.93	75.40	65.09	63.43	66.62	66.25
MM-DFN	44.44	44.44	77.55	80.00	71.35	66.99	75.88	70.88	74.25	76.42	58.27	61.67	67.84	67.85
TS-GCL	71.20	70.00	81.30	81.70	67.40	64.20	60.50	61.40	74.60	76.50	62.00	64.00	70.30	70.20
MultiEMO	53.80	56.29	83.33	83.50	**75.60**	70.11	68.29	67.07	79.70	75.79	64.82	70.35	72.29	71.69
SDT	55.06	57.62	80.58	80.08	65.73	69.14	67.88	66.87	82.50	73.47	66.58	67.53	70.54	69.95
GraphCFC	41.52	45.08	**87.12**	**84.84**	65.19	63.27	68.31	70.82	77.16	75.85	62.86	63.19	68.39	68.02
RL-EMO	40.28	47.15	79.18	81.17	69.79	66.67	74.12	64.28	78.60	76.18	58.01	61.82	69.16	68.20
DEDNet	56.32	64.07	81.15	80.98	73.92	74.97	67.37	71.11	84.38	77.84	72.99	69.68	74.47	73.79
DER-GCN	60.70	58.80	75.90	79.80	66.50	61.50	71.30	72.10	71.10	73.30	66.10	67.80	69.70	69.40
MERC-PLTAF	**75.60**	**75.30**	74.20	80.00	71.10	71.40	**75.90**	**74.70**	59.70	54.70	**73.50**	**72.20**	72.70	71.40
AMB-DSGDN（我们的方法）	60.23	66.25	84.58	81.36	74.14	**76.20**	70.86	71.88	85.87	**81.34**	73.24	72.17	**76.09**	**75.64**

表3该表使用 F1分数评估所有模型在 MELD（七种情绪类别）数据集上的性能，并以加权准确率和加权 F1 为指标展示模型在三个数据集上的整体性能，其中最佳结果以粗体标示，次佳结果加下划线。

模型	MELD															
	中性		惊讶		fear		悲伤		joy		厌恶		愤怒		加权准确率	加权 F1
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1		
DialogueRNN	82.17	76.56	46.62	47.64	0.00	0.00	21.15	24.65	49.50	51.49	0.00	0.00	48.41	46.01	60.27	57.95
MMGCN	84.32	76.96	47.33	49.63	2.00	3.64	14.90	20.39	56.97	53.76	1.47	2.82	42.61	45.23	61.34	58.41
MM-DFN	79.06	75.80	53.02	50.42	0.00	0.00	17.79	23.72	59.20	55.48	0.00	0.00	50.43	48.27	60.96	58.72
TS-GCL	78.10	**80.60**	56.70	56.40	6.80	5.20	42.30	43.70	**68.30**	66.30	2.30	2.60	43.80	48.50	64.40	64.10
MultiEMO	76.44	79.42	55.92	58.12	25.71	21.18	51.05	41.60	62.23	63.07	**45.71**	31.07	55.28	53.37	65.45	65.77
SDT	75.99	79.65	**59.21**	58.78	24.44	23.16	**59.22**	39.23	64.40	62.76	40.00	**31.86**	52.27	54.44	66.00	65.92
GraphCFC	71.26	75.17	46.18	45.68	9.09	3.28	30.43	11.42	50.26	49.49	0.00	0.00	35.88	41.93	54.34	55.20
RL-EMO	**85.59**	79.57	58.72	59.03	14.00	16.09	18.27	27.64	60.45	63.53	16.18	20.18	**55.36**	52.84	65.63	63.47
DEDNet	76.51	79.97	56.77	58.90	21.88	17.07	50.38	39.30	64.97	62.63	40.54	28.57	53.65	**54.49**	65.52	65.88
MERC-PLTAF	82.90	80.50	59.10	59.10	24.00	**26.90**	55.30	**46.40**	63.40	**77.10**	17.60	26.10	50.30	53.90	**68.00**	52.80
AMB-DSGDN（我们的方法）	75.16	80.00	58.48	**59.30**	**30.00**	17.14	58.18	40.25	64.75	64.59	44.44	25.26	54.93	54.12	66.07	**66.18**

基准概率为 $q_{\text{base}} = 0.3$ ，执行概率为$p_{\text{exe}} = 0.5$。

6. 结果与分析

本节从多个角度系统评估所提出的AMB-DSGDN 模型，包括整体性能、消融研究、参数敏感性、鲁棒性和计算复杂度，以全面分析其有效性和实用性。

6.1. 主要结果

表2 和3 展示了所提出的模型 AMB-DSGDN 在 IEMOCAP 和 MELD 数据集上的性能，并与各种基线模型进行了比较。为确保实验公平，所有模型均采用相同的预处理和训练设置。

在 IEMOCAP 数据集上，AMB-DSGDN 的加权准确率达到 76.09%，加权 F1 达到 75.64%，相比第二优模型 DEDNet 分别提升了 1.62% 和 1.85%。尤其是在愤怒、兴奋和挫败情绪上，该模型展现出显著优势，这得益于差分注意力图卷积网络在细粒度语义关系建模方面的能力，以及自适应模态丢弃策略对模态不平衡的有效缓解，使模型能够更好地捕捉对话中的关键信情感信息。

在 MELD 数据集上，AMB-DSGDN 的加权准确率为 66.07%，加权F1分数为 66.18%，与排名第二的模型相比提升有限。我们认为这由多种因素造成。首先，如表1所示，MELD 的类别分布高度不平衡，若干情绪类别（如厌恶、恐惧）的样本稀少，限制了模型对低频类别的学习和泛化能力；其次，MELD 是多说话人对话数据集，其中音频和视觉信号更容易受到说话人切换、语音重叠、面部遮挡等因素的干扰，从而降低单模态质量并增加模态融合的难度；尽管如此，该模型在惊讶等类别上仍表现出明显优势，展现了其在多说话人环境中的适应能力。

总体而言，得益于更高的模态质量和更清晰的情感关联，AMB-DSGDN 在 IEMOCAP 上展现出显著优势；而在 MELD 上，受类别分布不均衡以及多说话人导致的模态质量波动影响，整体提升较为有限，但它

表4不同模态组合在 IEMOCAP 和 MELD 数据集上的性能比较。

模态设置	IEMOCAP		MELD	
	加权准确率	加权 F1	加权准确率	加权 F1
A	62.60	60.70	34.70	37.70
T	69.02	68.63	65.61	65.78
V	39.80	35.60	23.16	31.27
A+T	73.43	72.87	65.49	65.65
V+T	69.71	69.16	65.77	65.86
A+V+T	**76.09**	**75.64**	**66.07**	**66.18**

在惊讶等类别上仍具有优势，体现了模型的鲁棒性和适应性。

6.2. 消融研究

为全面验证模型中关键组件的有效性，我们开展了一系列消融实验，重点考察模态组合、窗口大小、差分注意力机制和自适应模态丢弃策略等因素。通过逐步移除或调整这些模块，并在 IEMOCAP 和 MELD 数据集上重新训练和评估，我们量化了各组件的贡献及其协同效应。

多模态组合： 在表 4中，我们考察了不同模态组合对模型性能的影响，以验证多模态融合的有效性。具体而言，我们评估了单模态、双模态和全模态设置下的性能，并比较了两个数据集上的结果。结果表明，在单模态配置中，文本模态通常表现最佳，体现了其在捕捉对话语义和上下文信息方面的优势；其次是音频模态，这可能是因为它对语音语调和节奏较为敏感。相比之下，视觉模态表现相对较弱，这可能是因为视觉线索（如面部表情）会受到视角和对话场景动态变化的限制，从而导致信息提取不稳定。此外，在双模态设置中，音频与文本的组合显著优于视觉与文本的组合，表明音频模态通过提供额外的声学线索，更好地补充文本信息，从而提高情感识别准确率。尽管视觉与文本的融合有所提升，但增益有限，这表明在某些对话语境中，视觉模态可能会引入噪声或冗余。最终，全模态融合（即结合音频、视觉和文本）在两个数据集上均取得最佳表现，验证了多模态协同的重要性。通过利用

同时利用多个信息源，模型能够更全面地捕捉情感动态，缓解单模态或双模态方法的局限性，从而提升整体鲁棒性和泛化能力。

窗口大小设置： 为探究上下文范围对对话情感识别任务中图卷积网络的影响，本实验系统地

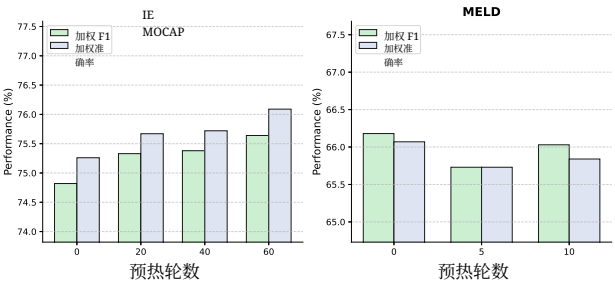


图3： IEMOCAP 和 MELD 数据集上不同窗口大小的实验结果。紫色线表示加权准确率，浅黄色柱表示加权 F1；左侧子图对应 IEMOCAP 数据集，右侧子图对应 MELD 数据集。注：窗口大小决定了图卷积网络的语义关联捕获范围，需要根据数据集特征进行调整。

表5语义图差分网络在 IEMOCAP 和 MELD 数据集上的消融研究。

变体	IEMOCAP		MELD	
	加权 F1	加权准确率	加权 F1	加权准确率
无关系子图 + DiffRGCN	68.00	68.82	65.07	65.10
无 DiffRGCN	73.44	73.56	65.62	64.80
完整模型	**75.64**	**76.09**	**66.07**	**66.18**

评估了不同窗口大小对模型性能的作用。图 3 显示，在 IEMOCAP 数据集上，随着窗口大小逐渐增加，模型的加权平均准确率和 F1 分数在中等窗口配置下达到最佳表现。这表明，适度的窗口能够充分覆盖语义相关的话语及其上下文依赖关系，从而增强图结构对情感关系的建模能力。窗口过小会限制对长程依赖信息的捕获，导致语义信息不足；窗口过大则可能引入无关特征，累积噪声，进而降低注意力权重计算和特征传播过程中的判别能力。在 MELD 数据集上，最佳表现则集中在相对较小的窗口区间。考虑到 MELD 的对话中说话人之间的互动更加频繁，且多轮结构更为复杂，较小的窗口能够更专注于局部语义和多模态信息交互，而过大的窗口则容易纳入不同说话人之间无关的情感特征，增加模态不一致性和分布噪声。这表明，不同数据集的对话结构、说话人分布和情感交互模式会显著影响窗口大小的最佳选择；在实际应用中，需要根据数据的统计特征调整这一超参数。

语义图差分网络： 为验证 DiffRGCN 和关系子图组件在整体模型中的作用，我们通过替换或移除相关模块，同时保持其余结构一致，设计了针对性的消融研究，以分析各组件对性能的独立

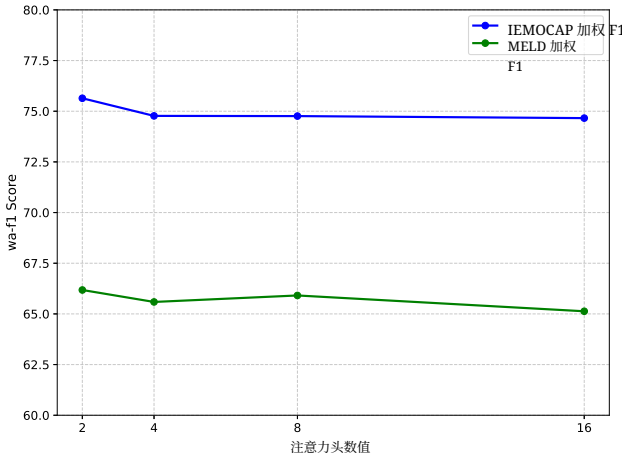


图4：注意力头数在 IEMOCAP 和 MELD 数据集上的敏感性分析结果。蓝线表示 IEMOCAP 数据集上的加权 F1，绿线表示 MELD 数据集上的加权 F1，横轴“注意力头数值”表示注意力头数，纵轴“加权平均F1分数”表示相应的加权平均F1分数。

贡献。需要注意的是，自适应模态平衡机制在这组消融研究中始终保持启用。具体而言，当仅移除 DiffRGCN 时，我们将差分关系图卷积替换为常规图卷积神经网络，在构建的关系子图上进行特征传播，其余模型架构保持不变；当同时移除关系子图和 DiffRGCN 时，模型不再执行图结构建模，而是在话语级多模态编码器之后，直接连接自适应模态平衡机制和后续分类模块。表 5 显示，同时移除关系子图和 DiffRGCN 会导致 IEMO-CAP 和 MELD 上的加权F1和准确率显著下降，验证了语义关系建模的重要性。进一步比较发现，仅用 GCN 替换 DiffRGCN 会导致性能下降，但仍优于完全不含图结构的变体，表明关系子图提供了基础语义关联，而 DiffRGCN 利用差分注意力刻画节点之间的细粒度语义差异，有效过滤冗余信息，并增强多模态情感依赖关系的表示。上述结果充分证明了语义图差分网络在多模态学习中的关键价值，以及其对模型性能和泛化能力的积极影响。

注意力头敏感性：在进一步的敏感性分析中，我们探究了差分注意力图卷积网络中的注意力头数量对模型性能的影响，以揭示其在多模态特征聚合中的作用。如图所示4，在 IEMOCAP 数据集上，模型在注意力头数量较少时表现最佳。

表6IEMOCAP 和 MELD 数据集上的模态平衡消融研究。

变体	IEMOCAP		MELD	
	加权 F1	加权准确率	加权 F1	加权准确率
不使用 模态丢弃 +关系子图 +DiffRGCN	67.83	68.78	64.87	64.80
不使用模态丢弃 +DiffRGCN	72.99	73.56	65.56	65.71
不使用模态丢弃	74.77	75.29	66.00	66.05
完整模型	**75.64**	**76.09**	**66.07**	**66.18**

这是因为少量注意力头能够有效捕捉语义多样性，同时保持特征表示紧凑。注意力头过少可能导致多模态信息整合不足，而注意力头过多则可能引入冗余特征或增加训练难度，从而降低模型的判别能力。在 MELD 数据集上，模型在注意力头数量较少的配置下同样表现出色，且随着注意力头数量增加，性能略有下降。这归因于 MELD 数据集的多说话人对话特点：在多说话人环境中，跨模态信号的整合难度更高；过多的注意力头可能引入不必要的噪声和信息冲突，而精简的注意力机制则能更集中地捕捉核心语义关联和情感动态，从而实现高效的信息整合。

自适应模态平衡：如表6所示，我们通过逐步移除核心模块，对自适应模态丢弃（MD）策略与语义图组件（关系子图和DiffRGCN）的协同效应进行了消融分析。在这组实验中，自适应模态平衡机制默认被禁用，网络其余结构与实验设置均与语义图差分网络小节保持一致。具体而言，当同时移除模态丢弃机制、关系子图和DiffRGCN时，模型不再执行模态调节或图结构建模，而是直接将话语级多模态表示输入分类模块；当移除模态丢弃机制和DiffRGCN时，模型仅保留关系子图用于特征聚合；当仅移除模态丢弃机制时，模型保留完整的语义图结构，但各模态以固定权重参与融合。实验结果表明，同时移除这三者会导致 IEMOCAP和MELD上的加权F1和准确率下降幅度最大，表明模态调节与语义关系建模之间存在显著互补性。进一步比较发现，未采用模态丢弃机制的模型在多模态特征融合中仍会出现性能下降，说明该机制能够动态调整各模态的参与程度，抑制主导模态的过度影响，从而在噪声干扰下提高情绪识别准确率。综上所述，自适应模态丢弃策略与语义图组件

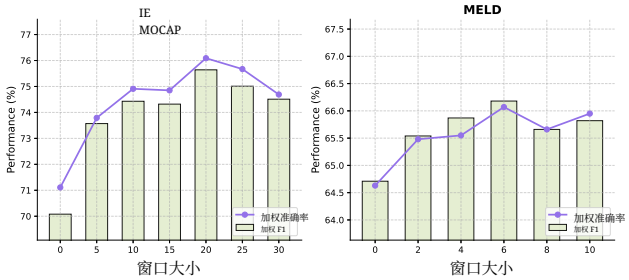


图 5： 预热阶段的性能动态。横轴表示预热期（即逐步启用模态丢弃机制前的训练轮数），纵轴表示模型性能指标；不同颜色的曲线分别对应模型在 IEMOCAP 和 MELD 数据集上的加权平均F1分数和准确率，展示了不同预热期对模型性能的影响。

在多模态对话情感识别中具有显著的互补效应，协同增强了模型捕捉情感动态依赖的能力以及对噪声干扰的鲁棒性。

预热期性能动态：为验证自适应模态丢弃策略的有效性，我们考察了预热期（即模态丢弃机制逐步激活前的训练轮数）对模型性能的影响，以揭示其在训练早期对模型稳定性与优化效果的作用。如图5所示，在 IEMOCAP 数据集上，随着预热期延长，模型性能先快速上升，约在 60 个训练轮数时达到峰值，随后保持稳定。这表明适度的预热有助于模型逐步适应模态不平衡问题，减少早期训练中的噪声干扰，从而提高对情感动态特征的捕捉准确率。若预热期过短，模型可能过早进入模态丢弃阶段，导致对各模态特征内部表征及跨模态关系的学习不足，进而造成性能波动或收敛不稳定。对于 MELD 数据集，在无预热时模型的初始性能较高，随着预热轮数增加，性能缓慢下降，最优性能出现在较短的预热期配置中。由于多说话人对话固有的复杂性，跨模态信号的整合难度相应增加；过长的预热期可能使模型过度依赖部分模态特征，而未能充分关注核心情感线索，从而引入冗余信息干扰。相反，较短的预热期可以快速激活模态丢弃机制，使模型更有效地识别并削弱冗余特征，从而高效整合跨模态信息，提高训练收敛速度并稳定性能。

表 7模态丢弃概率参数的敏感性分析结果。

IEMOCAP			
q_{base}	$p_{exe} = 0.3$	$p_{exe} = 0.5$	$p_{exe} = 0.7$
0.1	74.88	75.62	75.2
0.3	75.37	**75.64**	75.06
0.5	75.07	74.96	75.12
MELD			
q_{base}	$p_{exe} = 0.1$	$p_{exe} = 0.2$	$p_{exe} = 0.3$
0.1	65.91	65.8	65.78
0.2	65.80	**66.18**	65.81
0.3	65.87	65.87	65.72

丢弃概率设置：在自适应模态丢弃策略的参数敏感性分析中，我们通过在两个数据集上进行网格搜索，系统探究了基础丢弃概率 q_base 与执行概率 p_exe 的不同组合对模型性能的影响，以评估其优化效果。如表7所示，在 IEMOCAP 数据集上，模型性能在较低的基础丢弃概率配置下逐步提升，并在中等基础丢弃概率（如 0.3）与中等执行概率（如 0.5）的组合下达到峰值，随后趋于稳定。这表明适度丢弃能够平衡噪声抑制与关键信息保留；过低的丢弃可能导致噪声特征干扰融合结果，而过高的丢弃则会破坏语义表示并增加训练不稳定性。同样，在 MELD 数据集上，模型在较低基础丢弃概率下具有较高的起始性能，最优配置集中在相对较低的基础丢弃概率（如 0.2）与中等执行概率区间，这与该数据集的多说话人、丰富对话内容及多样情感特征密切相关。保守的丢弃策略能够有效保留各模态的核心信息，减少不同说话人之间信号差异所造成的干扰，同时平衡噪声抑制，使模型在复杂多样的对话环境中更加稳健。

6.3. 处理长序列对话

本节专门评估模型在长序列对话场景中的能力，重点关注其捕捉动态上下文依赖关系的能力。实验基于 IEMOCAP 数据集的长对话子集，选取包含 20 至 50 个话语的对话样本，以模拟真实交互场景中复杂且不断演变的时间依赖关系。使用加权 F1 分数和准确率评估模型性能，并与多种基线方法进行比较。

图中所示结果6 表明，所提出的方法在长序列对话子集上的表现显著优于基线模型，证明了其在建模长程上下文依赖方面的有效性。

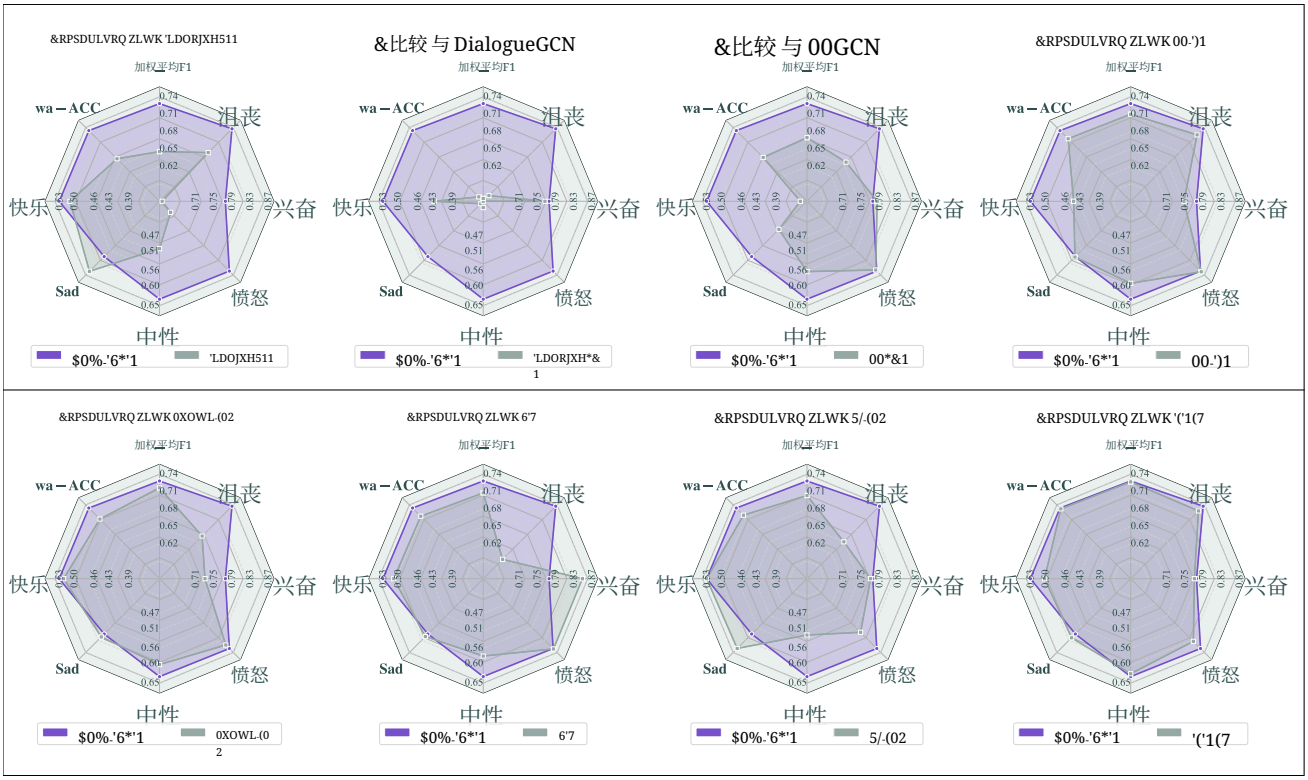


图6：所提方法与其他基线方法在IEMOCAP 数据集长序列对话子集上的性能对比，直观展示了所提方法在实验中的性能优势。

与基线相比，我们的模型即使在更长的对话序列上也能保持稳定的性能，没有明显下降，表明其捕获长期依赖的能力更强。

进一步分析表明，这一优势主要源于两方面设计。首先，动态跨模态贡献平衡机制根据对话进程实时调整不同模态的参与程度，避免单一模态主导并在长序列中造成信息偏差。其次，差分关系图建模显式捕捉节点之间的细粒度差异，有效缓解长序列任务中常见的上下文遗忘问题。在对话中期，基线方法常因累积噪声和模态偏移而出现性能波动，而所提方法通过自适应调整模态权重，保持稳定的情感识别。

在对话后期，我们的方法更有效地利用累积的上下文信息，抑制早期噪声对当前预测的影响，并减少误差传播。这表明该模型不仅能够捕捉局部情感变化，还能有效建模全局情感演化趋势。

从情感依赖演化的角度来看，长对话中的情绪呈现出明显的非线性和阶段性特征。在早期阶段，模型主要依靠局部模态特征进行判断。随着对话的推进，说话人之间的情感交互逐渐增强，差分机制捕捉正负子空间中的变化。在后期阶段，模型通过动态注意力强化关键上下文信息，有效整合整个序列中的信息。这一过程确保了对话中情感轨迹的全面建模，进一步验证了所提方法在长序列对话任务中的鲁棒性和适应性。

6.4. 计算开销增量分析

为全面评估所提出的差分图卷积网络和自适应模态平衡模块对计算的影响，我们对完整模型在 IEMOCAP 和 MELD 数据集上的推理时间、批处理时间、吞吐量及每秒浮点运算次数（浮点运算次数）进行了比较分析，结果如表 8 所示。

对于差分图卷积网络，与传统的图注意力网络相比，额外的计算开销主要来自引入差分注意力机制以及关系嵌入计算。通过对多个注意力分布之间的差异进行建模，该机制能够有效突出与上下文相关的依赖关系，从而增强对说话人内部情感连续性和说话人间情感交互的建模。由于每个节点都需要计算多组注意力权重

表8DiffRGCN 与标准图注意力网络的计算开销分析，以及自适应模态平衡模块的计算开销分析。

模块	数据集	推理时间（毫秒）		吞吐量（样本/秒）		批处理时间 (ms)		每秒浮点运算次数 (T)	
		基线	Δ (变化)	基线	Δ (变化)	基线	Δ (变化)	基线	Δ (变化)
DiffRGCN	IEMOCAP	0.0769	+0.0090 (+11.70%)	13169.51	−1693.29 (−12.86%)	62.14	+10.7027 (+17.22%)	0.0805	−0.0157 (−19.50%)
	MELD	0.4271	+0.1026 (+24.03%)	2384.02	−404.02 (−16.95%)	60.33	+11.4914 (+19.05%)	0.0500	−0.00236 (−4.72%)
自适应模态平衡	IEMOCAP	0.0769	+0.0165 (+21.45%)	13169.51	−980.73 (−7.45%)	62.14	+6.52 (+10.50%)	0.0805	−0.0082 (−10.19%)
	MELD	0.4271	+0.0190 (+4.45%)	2384.02	−137.90 (−5.79%)	60.33	+2.1635 (+3.59%)	0.0500	−0.00384 (−7.68%)

并执行差分运算，计算复杂度会随着图规模和边的数量增加，导致推理时间和批处理时间适度增加，同时吞吐量略有下降。值得注意的是，这种开销主要局限于局部子图内的注意力计算，不会引入全局高阶运算，因此对话长度增加时，复杂度的增长相对温和。在长序列对话场景中，DiffRGCN 仍可通过基于窗口的子图建模有效限制计算范围，从而保持良好的可扩展性。

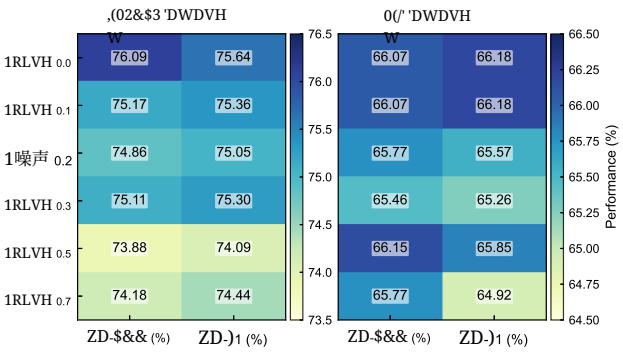


图 7： 模型在 IEMOCAP 和 MELD 数据集上的全模态噪声鲁棒性测试结果。图中的每个单元格展示了模型在不同噪声强度下的加权准确率和加权 F1。噪声强度范围为 0 至 0.7，表示测试时添加到文本、视觉和音频特征上的高斯噪声标准差倍数。颜色编码表示性能水平，颜色越深代表性能越高；右侧的颜色条提供相应百分比数值的参考。该热力图直观地展示了模型在不同噪声条件下的稳定性和性能变化。

对于自适应模态平衡模块而言，其核心目标是通过自适应模态丢弃策略，在多模态融合过程中动态调节不同模态的相对贡献，从而减轻单一模态在决策过程中的主导作用。该模块主要由轻量级操作构成，包括概率估计、随机采样和特征缩放，不会引入额外的复杂网络结构或大规模矩阵运算，因此对整体模型施加的计算负担有限。在长序列场景下，其计算复杂度大致随序列长度呈线性增长，使模型能够在对话长度增加时保持稳定的计算效率。该模块在保持计算效率的同时动态调整模态贡献，进一步增强了模型的

在复杂对话环境中的泛化能力和鲁棒性。

6.5. 多模态噪声鲁棒性分析：

为评估模型在复杂噪声条件下的稳定性和泛化能力，我们在 IEMOCAP 和 MELD 数据集上进行了全模态噪声干扰实验。测试期间，我们向文本、视觉和音频特征中注入了不同强度的高斯噪声，并按整体特征标准差对噪声幅度进行归一化，以模拟感知误差、环境干扰和模态质量波动。实验涵盖多个等级，从无噪声基线到高强度噪声（0.1–0.7），并采用加权准确率和加权平均 F1 分数作为评估指标。

如图 7所示，随着噪声强度增加，模型性能平稳下降，而非剧烈波动，表明其具有较强的鲁棒性。在 IEMOCAP 上，当全模态噪声达到 0.3 时，加权准确率和平均 F1 分数相较于基线仅略有下降；即使噪声达到 0.7，性能下降仍处于可控范围内。在 MELD 上，即使噪声水平达到 0.5–0.7，加权准确率和平均 F1 分数仍保持在 65% 左右，仅略低于基线，表明模型具有良好的噪声抗性。

进一步分析表明，DiffGCN 能够持续捕捉跨模态以及说话人内部和说话人之间的动态情感依赖关系。其差分注意力

表9模型参数和运行时间比较

方法	参数（M）	IEMOCAP运行时间（秒）	MELD运行时间（秒）
DialogueGCN	12.92	58.1	127.5
RGAT	15.28	68.5	146.3
LR-GCN	15.77	87.7	142.3
MMGCN	0.46	93.7	75.3
DER-GCN	78.59	125.5	189.7
AMB-DSGDN（我们的方法）	13.13	179.95	146.4

机制能够有效减少冗余的模态间干扰，在噪声条件下保留判别能力。同时，自适应模态平衡机制会在高噪声条件下动态调整模态权重，防止主导模态压制融合过程，同时保留次要模态中的有用信息，从而在全模态噪声环境下维持模型输出的一致性和可靠性。

6.6. 复杂度分析

表9 总结了各对比模型在参数量和运行时间方面的实验结果。从实验数据可以看出，尽管我们的模型参数规模与部分图卷积模型相近，但其推理时间略长，这与引入差分关系图卷积网络和自适应模态平衡模块直接相关。实验分析表明，差分关系图卷积网络会执行多次注意力计算，并为每个节点整合关系嵌入，从而增强模型捕捉说话人内部及说话人之间依赖关系的能力，但也带来了额外的计算开销。自适应模态平衡模块则会动态调整每个批次中各模态的贡献，确保弱模态信息得到有效利用；实验表明，它对整体推理时间的额外影响相对有限。尽管计算成本有所增加，但与性能提升相比，这种开销仍在可接受范围内。实验结果表明，该模型在高准确率、计算复杂度和多模态信息融合之间实现了良好平衡。

6.7. 极端不平衡下的自适应模态平衡

基于表 4 中的单模态性能差异，我们在 IEMOCAP 和 MELD 数据集上进行了极端模态不平衡实验，将文本、视觉和音频模态的融合权重设为 0.8，分别为 0.1 和 0.1。图8 的结果表明，尽管文本模态占主导地位，模型仍能保持较高的整体准确率和加权F1分数。对混淆矩阵的进一步分析显示，视觉和音频模态在识别“愤怒”和“兴奋”等关键情绪时仍有显著贡献，有效减轻了由文本模态偏差导致的误分类。

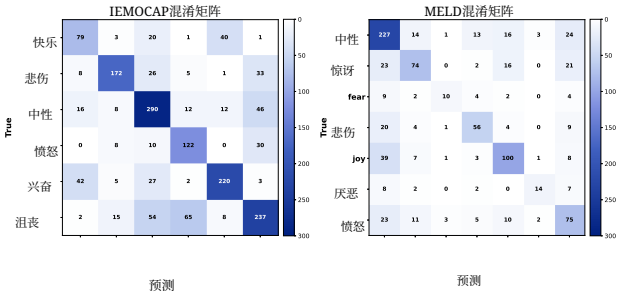


图8： 该图展示了极端模态不平衡实验下 IEMOCAP（左）和 MELD（右）数据集的情感识别混淆矩阵：左侧矩阵对应 IEMOCAP 的 6 个情绪类别，右侧矩阵对应 MELD 的 7 个情绪类别。“True”轴表示真实情感标签，“Predicted”轴表示模型的预测标签；单元格中的数值表示相应类别的样本数量。矩阵对角线上的深色区块反映了大多数情绪较高的识别准确率，而对角线之外的浅色区块则反映了不同情绪之间的误分类情况。

这表明，模型能够根据各模态的相对有效性自适应地调节信息流，防止主导模态过度影响融合结果。结合差分图注意力机制，这种调节还能够抑制共享噪声，确保不同模态在各个阶段都能做出更加均衡的贡献。总体而言，实验验证了自适应模态平衡机制在增强模型鲁棒性以及提升其对极端模态不平衡情况下动态多模态情绪变化的敏感性方面的有效性。

7. 结论

本研究提出了用于多模态对话情感识别的 AMB-DSGDN。该模型通过基于实时性能评估的自适应模态丢弃机制，有效缓解了模态贡献不平衡问题。同时，模型为每种模态分别构建说话人间和说话人内关系子图，并引入 DiffRGCN，通过正负子空间投影和差分注意力操作捕捉情感依赖关系的动态演化。实验结果表明，在 IEMOCAP 和 MELD 数据集上，AMB-DSGDN 的表现显著优于现有基线方法，在整体性能和特定情绪类别上均取得了稳定提升。

然而，由于图结构建模带来的计算复杂度和开销，所提出模型在处理超长对话序列以及在资源受限环境中运行时，推理效率仍然有限。尤其是，

要将模型部署在边缘设备上，必须妥善应对模型大小、图注意力机制的计算成本以及多模态特征的实时处理等挑战。未来的工作将着重探索轻量级图注意力设计、子图剪枝和参数共享策略，以提升计算效率；同时，还将研究模型压缩、知识蒸馏和硬件感知加速技术，从而提升模型在边缘设备上的实时推理性能。预计这些工作将进一步提升所提框架在实时人机交互和情感分析场景中的适用性。

CRediT作者贡献声明

Yunsheng Wang：概念构思、方法设计、调查研究、数据整理、撰写初稿。**YuntaoShou**：监督、调查研究与审阅。**Yilong Tan**：监督、调查研究、撰写与编辑审阅。**TaoMeng**：监督、调查研究、撰写与编辑审阅。**WeiAi**：监督、调查研究、撰写与编辑审阅。**KeqinLi**：监督、调查研究、撰写与编辑审阅。

利益冲突声明

作者声明，他们不存在已知的可能影响本文所报告研究工作的财务利益冲突或个人关系。

数据可用性

数据可应要求提供。

致谢

作者谨向匿名审稿人和副编辑致以诚挚谢意，感谢他们认真细致的工作和富有见地的建议，这些意见极大地改进了本文。本研究得到国家自然科学基金（项目编号：69189338）、中国湖南省优秀青年人才（项目编号：22B0275）以及复杂网络局部社区结构检测算法研究项目（项目编号：2020YJ009）的资助。

参考文献

[1] C. Wu、Y. Cai、Y. Liu、P. Zhu、Y. Xue、Z. Gong、J. Hirschberg 和 B. Ma，《对话中的多模态情感识别：方法、趋势、挑战与前景综述》，*arXiv* 预印本 *arXiv:2505.20511*，2025。[2] S. Poria、D. Hazarika、N. Majumder、G. Naik、E. Cambria 和 R. Mihalcea，《MELD：用于对话情感识别的多模态多方数据集》，*arXiv* 预印本 *arXiv:1810.02508*，2018。[3] W. Ai、F. Zhang、Y. Shou、T. Meng、H. Chen 和 K. Li，《从图谱视角重新审视对话中的多模态情感识别》，载于*AAAI*人工智能会议论文集，第39卷，第11418–11426页，2025。

[4] Y. Shou、T. Meng、W. Ai、F. Zhang、N. Yin 和 K. Li，“基于信息瓶颈的对抗对齐与图融合：用于对话中的多模态情感识别”，信息融合，第 112 卷，第 102590 页，2024 年。[5] Y. Wu、S. Zhang 和 P. Li，“基于提示学习与文本-音频融合特征的对话多模态情绪识别”，科学报告，第 15 卷，第 1 期，第 8855 页，2025 年。[6] G. Tu、F. Xiong、B. Liang、H. Wang、X. Zeng 和 R. Xu，“对话中的多模态情感识别校准”，收录于第 *32* 届 *ACM* 国际多媒体会议论文集，第 9621–9630 页，2024 年。[7] Y. Shou、H. Liu、X. Cao、D. Meng 和 B. Dong，“一种基于低秩匹配注意力的对话情感识别跨模态特征融合方法”，*IEEE*情感计算汇刊，第 16 卷，第 2 期，第 1177–1189 页，2024 年。[8] T. Meng、Y. Shou、W. Ai、N. Yin 和 K. Li，“对话中的多模态情感识别深度不平衡学习”，*IEEE*人工智能汇刊，第 5 卷，第 6487 页，2024 年 12 月，第 6472 页。[9] Y. Sun 和 T. Zhou，“DialogueMLLM：通过指令微调大语言模型推动对话多模态情感识别”，*IEEE Access*，第 13 卷，第 1–12 页，2025 年。[10] Y. Shou、T. Meng、W. Ai、F. Fu、N. Yin 和 K. Li，“基于深度学习的多模态对话情绪识别综合综述”，*ACM*信息系统汇刊，第 44 卷，第 2 期，第 1–48 页，2026 年。[11] Y. Shou、T. Meng、W. Ai、S. Yang 和 K. Li，“基于图卷积神经网络和依存句法分析的对话情感识别研究”，神经计算，第 501 卷，第 629–639 页，2022 年。[12] Y. Fu、X. Yan、W. Chen 和 J. Zhang，“用于对话情感识别的特征增强型多模态交互模型”，知识型系统，第 309 卷，第 112876 页，2025 年。[13] Y. Shou、X. Cao、H. Liu 和 D. Meng，“用于年龄估计的掩码对比图表示学习”，模式识别，第 158 卷，第 110974 页，2025 年。[14] T. Meng、Y. Shou、W. Ai、J. Du、H. Liu 和 K. Li，“基于异构图的对话情感识别多消息传递框架”，神经计算，第 569 卷，第 127109 页，2024 年。[15] H. Mehrez 和 S. A. Selouani，“连续情感空间中的对话系统多模态情绪识别”，收录于第 *6* 届智能技术生物工程国际会议（*BioSMART*）论文集，第 1–4 页，2025 年。[16] Y. Shou、W. Ai、T. Meng 和 K. Li，“图扩散模型：方法与应用综述”，计算机科学评论，第 59 卷，第 100854 页，2026 年。[17] Y. Shou、W. Ai、J. Du、T. Meng、H. Liu 和 N. Yin，“用于对话多模态情感识别的高效长距离潜在关系感知图神经网络”，*arXiv* 预印本 *arXiv:2407.00119*，2024 年。[18] H. Lian、C. Lu、S. Li、Y. Zhao、C. Tang 和 Y. Zong，“基于深度学习的多模态情感识别综述：语音、文本与人脸”，熵，第 25 卷，第 10 期，第 1440 页，2023 年。[19] Y. Shou、X. Cao 和 D. Meng，“Spegcl：无需正样本的自监督图谱对比学习”，*IEEE*神经网络与学习系统汇刊，2025 年。[20] W. Ai、Y. Tan、Y. Shou、T. Meng、H. Chen、Z. He 和 K. Li，“范式转变：用于多模态虚假新闻检测的大型视觉语言模型综述”，计算机科学评论，第 60 卷，第 100893 页，2026 年。[21] S. Zhao、J. Ren 和 X. Zhou，“用于对话多模态情感识别的跨模态门控特征增强”，科学报告，第 15 卷，第 1 期，第 30004 页，2025 年。[22] H. Zhao、Y. Gao、H. Chen、B. Li、G. Ye 和 Z. Zhang，“通过上下文过滤和多频率图传播增强对话中的多模态情感识别”，收录于 *ICASSP 2025—2025 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 1–5 页，IEEE，2025 年。

[23] Z. Cheng、Z.-Q. Cheng、J.-Y. He、K. Wang、Y. Lin、Z. Lian、X. Peng 和 A. Hauptmann，“Emotion-llama：通过指令微调实现多模态情绪识别与推理”，《神经信息处理系统进展》，第 37 卷，第 110805–110853 页，2024 年。[24] A. Farhadipour、H. Ranjbar、M. Chapariniya、T. Vukovic、S. Ebling 和 V. Dellwo，“多方对话场景中的多模态情绪识别与情感分析”，*arXiv* 预印本 *arXiv:2503.06805*，2025 年。[25] Y. Shou、H. Lan 和 X. Cao，“结合对抗式跨视图重构与信息瓶颈的对比图表示学习”，神经网络，第 184 卷，第 107094 页，2025 年。[26] Y. Shou、T. Meng、W. Ai 和 K. Li，“基于广义状态空间模型与概率引导融合的多模态情绪学习再探”，收录于欧洲机器学习与数据库知识发现联合会议，第 509–525 页，施普林格，2025 年。[27] Y. Shou、T. Meng、W. Ai 和 K. Li，“用于对话多模态情感识别的动态图神经常微分方程网络”，收录于第 *31* 届国际计算语言学会议论文集，第 256–268 页，2025 年。[28] Q. Fan、H. Zuo、R. Liu、Z. Lian 和 G. Gao，“不完整数据场景下用于多模态情绪识别的抗噪联合表示学习”，收录于第 *2* 届多模态与负责任情感计算国际研讨会论文集，第 116–124 页，2024 年。[29] Y. Shou、J. Yao、T. Meng、W. Ai、C. Chen 和 K. Li，“GSDNet：从图谱视角重新审视不完整多模态扩散情绪识别”，收录于第三十四届国际人工智能联合会议论文集，*IJCAI-25*。国际人工智能联合会议组织，第 6182–6190 页，2025 年。[30] Y. Shou、W. Ai、T. Meng、F. Zhang 和 K. Li，“GraphUNet：图构建强大的遥感分割编码器”，收录于*2023 IEEE*第*29*届国际并行与分布式系统会议（*ICPADS*），第 2734–2737 页，IEEE，2023 年。[31] Y. Shou、T. Meng、W. Ai 和 K. Li，“用于对话多模态情感识别的动态图神经常微分方程网络”，*arXiv* 预印本 *arXiv:2412.02935*，2024 年。[32] Y. Shou、X. Cao、P. Yan、Q. Hui、Q. Zhao 和 D. Meng，“基于双分支编码器和双层对齐的图域适应，用于基于全切片图像的生存预测”，收录于*IEEE/CVF*国际计算机视觉会议论文集，第 19925–19935 页，2025 年。[33] Y. Shou、T. Meng、W. Ai 和 K. Li，“多模态大语言模型与多模态情感识别及推理：综述”，*arXiv* 预印本 *arXiv:2509.24322*，2025 年。[34] N. Majumder、S. Poria、D. Hazarika、R. Mihalcea、A. Gelbukh 和 E. Cambria，“DialogueRNN：用于对话情绪检测的注意力循环神经网络”，收录于*AAAI*人工智能会议论文集，第 33 卷，第 6818–6825 页，2019 年。[35] Y. Shou、T. Meng、W. Ai、N. Yin 和 K. Li，“CILF-CIAE：用于校正逆向年龄估计的 CLIP驱动的图像—语言融合”，神经网络，第 108518 页，2025 年。[36] D. Ghosal、N. Majumder、S. Poria、N. Chhaya 和 A. Gelbukh，“DialogueGCN：用于对话情感识别的图卷积神经网络”，*arXiv* 预印本 *arXiv:1908.11540*，2019 年。[37] L. Guo、Y. Song 和 S. Ding，“结合跨模态注意力的说话人感知认知网络，用于对话多模态情感识别”，知识型系统，第 296 卷，第 111969 页，2024 年。[38] Y. Liu、H. Zhang、Y. Zhan、Z. Chen、G. Yin、L. Wei 和 Z. Chen，“用于情感识别的抗噪多模态变换器”，《国际计算机视觉杂志》，第 1–21 页，2024 年。[39] Z. Lian 、 H. Chen 、 L. Chen 、 H. Sun 、 L. Sun、Y. Ren、Z. Cheng、B. Liu、R. Liu、X. Peng、等，“AffectGPT：用于多模态大语言模型情绪理解的新数据集、模型与基准”，*arXiv* 预印本 *arXiv:2501.16566*，2025 年。[40] T. Sun、Y. Wei、J. Ni、Z. Liu、X. Song、Y. Wang 和 L. Nie，“通过层次化知识蒸馏实现多模态情绪识别”，*IEEE*多媒体汇刊，第 26 卷，第 9036–9046 页，2024 年。

[41] Z. Lian, L. Sun, Y. Ren, H. Gu, H. Sun, L. Chen, B. Liu和J. Tao，“Merbench：多模态情感识别统一评测基准”，*arXiv*预印本 *arXiv:2401.03429*，2024年。[42] G. R. K. Reddy、A. D. Bhavani 和V. K. Odugu，“基于优化循环神经网络的脑部情感识别技术”，*Multimedia Tools and Applications*，第84卷，4674，2025年8月，第4655页–。[43] W. Ai、Y. Shou、T. Meng和K. Li，“Der-gcn：用于多模态对话情感识别的对话与事件关系感知图卷积神经网络”，*IEEE Transactions on Neural Networks and Learning Systems*，第36卷，第3期，第4908–4921页，2024年。[44] D. Chandola、E. Altarawneh、M. Jenkin和M. Papagelis，“Serc-gcn：使用图卷积网络进行对话中的语音情感识别”，载于*ICASSP 2024-2024 IEEE International Confer- ence on Acoustics, Speech and Signal Processing (ICASSP)*，第76–80页，IEEE，2024年。[45] H. Kong、X. Lou和Z. Li，“基于GCN与LSTM结合的情感脑电识别与分类”，载于*2024IEEE International Conference on Medical Artificial Intelligence (MedAI)*，第294–301页，IEEE，2024年。[46] Y. Wang、W. Zhang、K. Liu、W. Wu、F. Hu、H. Yu和G. Wang，“通过关系子图交互实现多模态情感识别的动态情感依赖网络”，*IEEETransactions on Affective Computing*，2024年。[47] Y. Wei、R. Feng、Z. Wang和D. Hu，“通过样本级模态评估增强多模态协作”，载于*Proceedings of theIEEE/CVF Conference on Computer Vision andPattern Recognition*，第27338–27347页，2024年。[48] H. Wang、W. Yang、X. Zhong、J. Zhou、F. Liu和W. Zhang，“缓解多模态在线联邦学习中的模态数量与质量不平衡”，*arXiv*预印本 *arXiv:2508.11159*，2025年。[49] Z. Guo、T. Jin和Z. Zhao，“面向情感分析与情感识别的缺失模态多模态提示学习”，*arXiv*预印本 *arXiv:2407.05374*，2024年。[50] Z. Lian、H. Sun、L. Sun、Z. Wen、S. Zhang、S. Chen、H. Gu、J. Zhao、Z. Ma、X. Chen、等人，“MER 2024：半监督学习、噪声鲁棒性与开放词汇多模态情感识别”，载于第二届 *International Workshop on Multimodal an d*

负责任情感计算，第 41–48 页，2024 年。[51] J. Chen、D. Yang、Y. Jiang、M. Li、J. Wei、X. Hou 和 L. Zhang，《聚焦效率：层归一化作为微调医学视觉语言预训练模型的催化剂》，*arXiv* 预印本 *arXiv:2404 16385*，2024 年。。[52] Y. Liu、M. Ott、N. Goyal、J. Du、M. Joshi、D. Chen、O. Levy、M. Lewis、L. Zettlemoyer 和 V. Stoyanov，《RoBERTa：一种稳健优化的 BERT 预训练方法》，*arXiv* 预印本 *arXiv:1907.11692*，2019 年。[53] G. Huang、Z. Liu、L. Van Der Maaten 和 K. Q. Weinberger，《密集连接卷积网络》，载于论文集 *o f*

*IEEE*计算机视觉与模式识别会议，第4700–4708页，2017年。[54] F. Eyben、M. Wöllmer 和 B. Schuller，“OpenSmile：慕尼黑通用快速开源音频特征提取器”，载于第*18*届*ACM*多媒体国际会议论文集，第1459–1462页，2010年。[55] C. Busso、M. Bulut、C.-C. Lee、A. Kazemzadeh、E. Mower、S. Kim、J. N. Chang、S. Lee 和 S. S. Narayanan，“IEMOCAP：交互式情感双人动作捕捉数据库”，语言资源与

评估，第 42 卷，第 4 期，第 335–359 页，2008 年。[56] J. Hu、Y. Liu、J. Zhao 和 Q. Jin，《Mmgcn：通过深度图卷积网络进行多模态融合的对话情感识别》，*arXiv* 预印本 *arXiv:2107.06779*，2021 年。[57] D. Hu、X. Hou、L. Wei、L. Jiang 和 Y. Mo，《Mm-dfn：用于对话中的情感识别的多模态动态融合网络》，载于*ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech andSignal Processing (ICASSP)*，第 7037–7041 页，IEEE，2022 年。[58] A. Joshi、A. Bhat、A. Jain、A. V. Singh 和 A. Modi，《COGMEN：基于上下文化图神经网络的多模态情感识别》，*arXiv* 预印本 *arXiv:2205.02455*，2022 年。

[59] T. Shi 和 S.-L. Huang，《MultiEMO：一种用于对话情感识别的基于注意力且感知相关性的多模态融合框架》，载于计算语言学协会第 *61* 届年会论文集（第 *1* 卷：长篇论文），第 14752–14766 页，2023 年。[60] H. Ma、J. Wang、H. Lin、B. Zhang、Y. Zhang 和 B. Xu，《一种采用自蒸馏的基于Transformer的对话多模态情感识别模型》，*IEEE*多媒体汇刊，第 26 卷，第 776–788 页，2024 年。[61] J. Li、X. Wang、G. Lv 和 Z. Zeng，《GraphCFC：一种基于有向图的跨模态特征补全方法，用于多模态对话情感识别》，*IEEE*多媒体汇刊，第 26 卷，第 77–89 页，2023 年。[62] C. Zhang、Y. Zhang 和 B. Cheng，《RL-EMO：一种用于多模态情感识别的强化学习框架》，载于*ICASSP 2024-2024IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*，第 10246–10250 页，IEEE，2024 年。