计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(6): 1478−1491, 2026

基于混合融合专家的多模态情感识别方法

陈海丰[1] 李 溪[1] 李 岩[2] 李 健[3] 何 浪[4] 蒋冬梅[5]

[1](陕西科技大学电子信息与人工智能学院 西安 710021)

[2](西北工业大学计算机学院 西安 710072)

[3](陕西科技大学电气与控制工程学院 西安 710021)

[4](西安邮电大学计算机学院 西安 710061)

[5](鹏城实验室 广东深圳 518055)

(chenhaifeng@sust.edu.cn)

Multimodal Emotion Recognition Method Based on Mixture of Fusion Experts

Chen Haifeng[1], Li Xi[1], Li Yan[2], Li Jian[3], He Lang[4], and Jiang Dongmei[5]

[1](*School of Electronic Information and Artificial Intelligence*, *Shaanxi University of Science & Technology*, *Xi'an* 710021)

[2](*School of Computer Science*, *Northwestern Polytechnical University*, *Xi'an* 710072)

[3](*School of Electrical and Control Engineering*, *Shaanxi University of Science & Technology*, *Xi'an* 710021)

[4](*School of Computer Science & Technology*, *Xi'an University of Posts and Telecommunications*, *Xi'an* 710061)

[5](*Pengcheng Laboratory*, *Shenzhen*, *Guangdong* 518055)

Abstract Multimodal emotion recognition aims to integrate data from multiple modalities for accurate inference of emotional states. Existing research either treats all modalities equally or adopts fixed fusion strategies based on a single modality, failing to adequately address the imbalance of modality contributions. To tackle this, we propose a dynamic multimodal fusion method MoMFE based on the mixture of experts (MoE) framework, incorporating an adaptive router module and a multimodal fusion expert module. This method dynamically evaluates the contribution of each modality and selects appropriate fusion strategies accordingly. The adaptive router mechanism dynamically assesses modality contributions through inter-modal correlation analysis and dynamic weighting to guide the selection of fusion experts, while an expert-guided loss function is integrated to further optimize the expert selection process. The multimodal fusion experts perform complementary fusion for different modality contribution combinations and introduce a shared expert to mitigate the loss of global information and reduce parameter redundancy. Comparative experiments on three benchmark datasets (MER2024, CMU-MOSEI, and CH-SIMS) for multimodal emotion recognition demonstrate that MoMFE method outperforms state-of-the-art (SOTA) multimodal fusion methods in core metrics, including binary emotion recognition accuracy (Acc-2) and $F1$ score. Notably, the method achieves an average improvement of approximately 2% on CH-SIMS dataset.

Key words emotion recognition; multimodal fusion; mixture of experts; adaptive router; cross-modal attention

收稿日期: 2025-07-20; **修回日期**: 2025-12-15

基金项目: 国家自然科学基金重点项目(62236006); 国家自然科学基金青年科学基金项目(62306172); 国家自然科学基金项目(62376215); 陕西省重点研发计划项目(2025CY-YBXM-191); 陕西省社会科学基金项目(2021K05)

This work was supported by the Key Program of the National Natural Science Foundation of China (62236006), the National Natural Science Foundation of China for Young Scientists (62306172), the National Natural Science Foundation of China (62376215), the Key Research and Development Program of Shaanxi Province (2025CY-YBXM-191), and the Social Science Foundation of Shaanxi Province (2021K05).

通信作者: 李健(lijianjsj@sust.edu.cn)

摘　要　多模态情感识别旨在整合不同模态数据以推断情感状态。现有研究通常平等对待各模态或依赖单一模态设计固定策略的融合方法，难以应对模态数据贡献度不均衡的问题。为此，提出一种动态的多模态混合融合方法（MoMFE），并基于混合专家模型框架设计了自适应路由模块和多模态融合专家模块，动态评估各模态的贡献度并选择相应的融合策略。其中，自适应路由机制通过模态间相关性分析动态评估模态贡献度，并为融合专家的选择提供指导，同时引入专家引导损失，进一步优化融合专家选择；多模态融合专家则针对不同模态贡献度组合进行互补融合，并通过引入共享专家以避免全局信息丢失和参数冗余。基于 MER2024, CMU-MOSEI, CH-SIMS 数据集上的多模态情感识别任务的对比实验表明，相比其他先进的多模态融合方法，MoMFE 方法在情感识别准确率（Acc-2）、*F*1 分数等核心评估指标上均取得优异的性能。

关键词　情感识别；多模态融合；混合专家；自适应路由；跨模态注意力

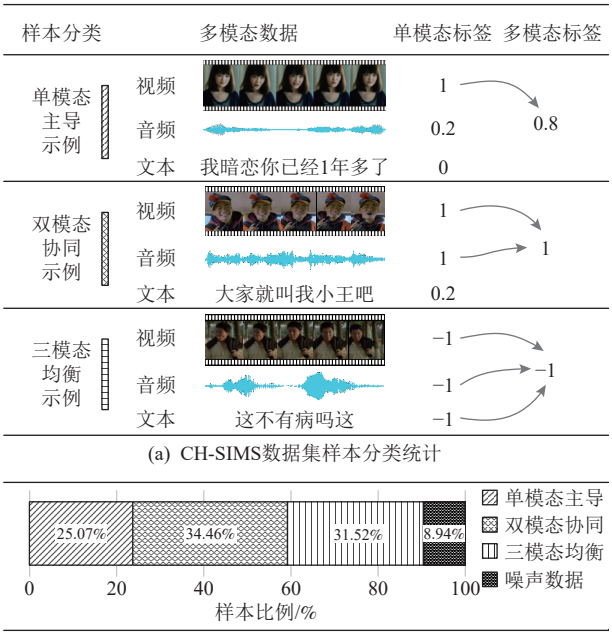
中图法分类号　TP391

DOI: 10.7544/issn1000-1239.202550544　　**CSTR:** 32373.14.issn1000-1239.202550544

情感识别通过文本、语音、图像等数据识别情感状态，在人机交互、舆情分析等领域广泛应用[1]。多模态情感识别凭借整合多种模态数据的优势日益受到重视[2-3]。而多模态情感识别的一大挑战是如何有效地融合不同模态传达的情感信息[4]，这主要源于人类复杂的情感表达方式，导致不同模态包含的情感信息量及其对情感识别任务的重要性不同。

现有研究主要通过单一固定的融合方式实现多模态融合。例如，Lian 等人[5] 通过计算模态间注意力分数作为权重进行加权融合；Luo 等人[6] 提出对情感语义表达一致的各模态特征表示进行加权连接；Liang[7] 和 Feng 等人[8] 分别利用遗传算法和情感知识指导衡量每个模态的贡献，从而选出主导模态，构建单一模态为主导的多模态融合。然而面对不平衡的模态贡献分布，单一的融合范式缺乏对模态间主导关系的进一步研究，动态选择多种融合策略的方法被引入到多模态融合中。申锦涛[9] 提出一种基于 GRU 门的多模态动态融合方法，选择是否以文本模态为主而进行加权融合，利用门控实现对文本主导和非文本模态主导的 2 种融合方法的选择。然而，在实际场景中，主导模态往往不是单一的，模态贡献度组合多样化。如图 1 呈现了 CH-SIMS 数据集的样本分类统计，该数据集同时包含单模态和多模态标签。经分析发现，数据集中的模态贡献度呈现出 3 种典型分布模式：单模态主导型（单一模态贡献显著）、双模态协同型（2 种模态共同主导）以及三模态均衡型（各模态贡献相当）。因此，需要依据不同模态贡献度分布动态选择对应的互补融合方法。

近年来，混合专家（mixture of experts, MoE）模型凭借其动态选择专家和数据处理的灵活性，在情感



(a) CH-SIMS数据集样本分类统计

(b) 分类示例

Fig. 1　Classification statistics of CH-SIMS dataset samples and classification examples

图 1　CH-SIMS 数据集样本分类统计以及分类示例

识别领域广泛应用。例如：Zhuang 等人[10] 利用 MoE 层的每个专家专注于单模态表示的不同方面，自主融合针对每个模态独特的多个局部表示；Zhao 等人[11] 则是将 MoE 作为多模态融合后的增强模块，通过加权专家组合来增强多模态表示中的情感信息细节。以上方法存在 2 个关键局限：一是多数工作仅将 MoE 用于单模态表示增强或后处理阶段，缺乏面向多模态融合策略的动态选择机制；二是所设计的专家结构多采用标准前馈网络，难以针对不同模态间的贡献分布进行专业化建模，从而限制了多样化融合需求的适应能力。为此，本文提出构建具备不同专业

化属性的互补融合专家。这一设计不仅使 MoE 框架有望成为实现动态融合的有效工具,更能通过组建多样化且专业化的互补融合专家、灵活选择适配的融合策略,进一步提升多模态特征的融合效果。

本文提出一种多模态混合融合专家(mixture of multimodal fusion experts, MoMFE)方法来实现自适应的多模态融合。MoMFE 基于混合专家模型框架,设计了自适应路由模块(adaptive router module, ARM)和多模态融合专家模块(multimodal fusion expert module, MFEM),实现依据模态贡献度分布动态选择融合方法。具体而言,在 ARM 中,首先计算模态间相互依赖系数,以此量化模态间相关性和信息共享程度,并为路由选择融合专家提供指导,从而增强路由对模态间主导关系的感知能力。接着引入专家引导损失来优化专家选择,确保选择的融合专家更符合模态主导关系,减少因信息不对称导致的路由决策偏差。在 MFEM 中,基于跨模态注意力机制计算模态间的加权特征进行互补融合,构建了多种适应不同模态主导关系的融合方法,为多模态特征融合提供了更灵活且适配的解决方案,以解决单一固定融合方式所造成的重要信息丢失问题。本文的主要贡献总结有 3 点:

1)提出一种基于混合专家模型的多模态融合方法,可依据各模态的贡献度动态选择适配的融合策略;

2)在路由机制中引入模态间相关性分析模块,通过动态评估模态贡献度,提升其对模态间相关性的感知能力,实现对融合专家有效的决策能力;

3)根据模态贡献度组合,利用跨模态注意力机制设计了对应的融合专家,最大化利用互补模态的有效情感信息。通过在 MER2024, CMU-MOSEI, CH-SIMS 这 3 个数据集上进行对比实验和消融实验,验证了本文提出方法在多模态情感识别任务中的有效性。

1 相关工作

1.1 多模态情感识别

情感识别研究最初主要围绕文本[12]、视觉[13]和音频[14]等单一模态展开。随着深度学习的发展,各模态均经历了从传统手工特征到深层特征的演进。其中文本模态从早期词袋模型,逐渐过渡到基于 GloVe[15]的预训练词向量,再到 BERT[16]等大规模预训练语言模型生成的上下文情感表征;视觉模态从人工设计的方向梯度直方图(histogram of oriented gradients, HOG)等图像特征,发展为通过卷积神经网络以及 Vision Transformer[17]等,逐步提升面部表情和肢体动作的识别精度;音频模态也从传统的梅尔频谱系数(Mel-frequency cepstral coefficients, MFCC)等声学特征,发展到借助 Hubert[18]和 Whisper[19]等语音预训练模型,更高效地提取与情感相关的深度特征。这些单模态领域的突破为多模态情感识别研究提供了强大的特征提取工具,为多模态情感识别中的特征融合与交互建模奠定了基础。

多模态融合是多模态情感识别任务中的关键步骤,早期融合和后期融合是 2 种经典的多模态融合策略。早期融合在特征提取阶段将不同模态的特征进行简单拼接,形成一个统一的特征表示。后期融合在每个模态独立进行特征提取和预测后,将各模态的预测结果进行融合。尽管以上方法在一定程度上实现了多模态信息的整合,由于情感表达在多模态数据上传递的信息量并不相同,过于简单的特征处理无法使多模态融合得到最有效的信息。最近许多工作致力于分析模态之间的信息相关性来解决模态不平衡问题。Wei 等人[20]将互信息与多模态情感识别结合,提出了一种层次互信息最大化框架,以减少与任务相关的信息丢失。Cai 等人[21]通过计算模态之间的互信息并利用互信息神经估计方法,最大化不同模态之间特征建模的相似性。此外,一些框架尝试根据不同模态对最终情感任务的贡献设计融合方法。Kumar 等人[22]提出一种条件门控机制,基于说话者的语言、语音和面部表情,选择性地学习不同模态的重要性权重。Majumder 等人[23]则通过使用门控网络自适应调节各个模态特征的重要性权重,从而实现多模态融合。上述方法在多模态融合过程中实现了对不同模态的差异化侧重,但其特征融合层并没有根据不平衡的模态而设计个性化的融合方法。因此,本文通过动态分析模态间相关性,构建了差异化的互补融合方案。

1.2 MoE 模型

MoE 是一种结合多个专家子模型的集成学习方法,最初由 Jacobs 等人[24]提出。MoE 的核心思想是希望将输入空间划分为多个类别,每个区域由一个专家模型负责处理,并通过路由机制动态选择最合适的专家进行后续任务。早期 MoE 主要应用于分类和回归任务,但随着深度学习的发展,MoE 逐渐被融入更复杂的神经网络架构中。近年来,MoE 在大规模预训练模型中的应用取得了显著进展。例如,Switch Transformer[25]是将 MoE 结构引入 Transformer[26]

模型的方案，通过稀疏激活机制，仅激活一小部分专家模型，显著提升了模型的参数效率和训练速度。随后，Gshard[27] 和 GlaM[28] 等模型进一步优化了 MoE 的分布式训练策略，解决了专家负载不均衡等问题。此外，研究人员还在 MoE 的动态路由、专家数量优化和路由机制设计等方面进行了深入探索[29]。

随着多模态数据在人工智能应用中的重要性日益凸显，MoE 架构已被广泛应用于各种多模态的下游任务中。例如，在多模态 Transformer 模型中，MoE 结构被用于动态选择不同模态的专家网络[30]，以增强模型对多模态数据的语义理解能力和泛化性能。然而，在情感识别任务中，现有的 MoE 架构在实现高质量的融合学习效果方面仍面临挑战。为了进一步研究针对性的模态特征融合专家，Lin 等人[31] 提出细粒度的专家融合，为每个模态单独构建视觉专家。Han 等人[32] 集成了 MoE 融合层与模态嵌入，并提供了多种路由器配置，以处理不同复杂度级别的多模态组合。尽管上述 MoE 模型考虑到多种模态数据的特性，但是缺乏对专家的针对性优化。Xiong 等人[33] 通过构建主模态与其他参考模态融合的专家模块，并引入模态重要性权重，在融合过程中选择性地关注来自特定模态的信息，这种方法在处理特定模态组合时表现出色。但在扩展到三模态或缺失模态时有局限性。因此，本文基于 MoE 框架，实现动态选择不同专业化的互补融合方法。

2 方法模型

2.1 总体架构

如图 2(a)所示，整体情感识别过程主要包括多模态特征提取、MoMFE 以及情感识别。其中 MoMFE 模块由 ARM 和 MFEM 组成，ARM 中引入相关性分析模块，MFEM 则包含多个不同的互补融合方法。整体工作流程如下：

输入数据包含视觉(v)、音频(a)和文本(t)这 3 种模态。首先，通过单模态编码器分别提取各模态的特征，得到 $\boldsymbol{f}_{\mathrm{v}} \in \mathbb{R}^{l_{\mathrm{v}} \times d_{\mathrm{v}}}$，$\boldsymbol{f}_{\mathrm{a}} \in \mathbb{R}^{l_{\mathrm{a}} \times d_{\mathrm{a}}}$，$\boldsymbol{f}_{\mathrm{t}} \in \mathbb{R}^{l_{\mathrm{t}} \times d_{\mathrm{t}}}$，其中 l_m 表示序列长度，d_m 表示特征向量的维度，$m \in \{\mathrm{v,a,t}\}$。随后，利用模态特定的多层感知机(multilayer perceptron, MLP)将多模态特征映射到同一维度 d。接着，将各模态特征输入 ARM，分析模态间的相关性并计算激活专家的概率 ***logits***，同时依据贡献度大小将多模态特征 $\boldsymbol{f}_{\mathrm{v}}$，$\boldsymbol{f}_{\mathrm{a}}$，$\boldsymbol{f}_{\mathrm{t}}$ 排序为 $\boldsymbol{f}_1$，$\boldsymbol{f}_2$，$\boldsymbol{f}_3$。之后，根据计算得到的 ***logits*** 来激活 MFEM 中最优的融合专家(Top-1)，对 $\boldsymbol{f}_1$，$\boldsymbol{f}_2$，$\boldsymbol{f}_3$ 进行互补融合。最终，融合特征 $\boldsymbol{F}_m \in \mathbb{R}^{l_m \times d_m}$ 被输入分类器，从而确定情感状态。

2.2 自适应路由

MoE 模型中的路由机制通常在词(token)级别进行专家选择，通过计算输入词与各专家之间的匹配分数来分配数据处理任务[34]。而针对情感识别任务，

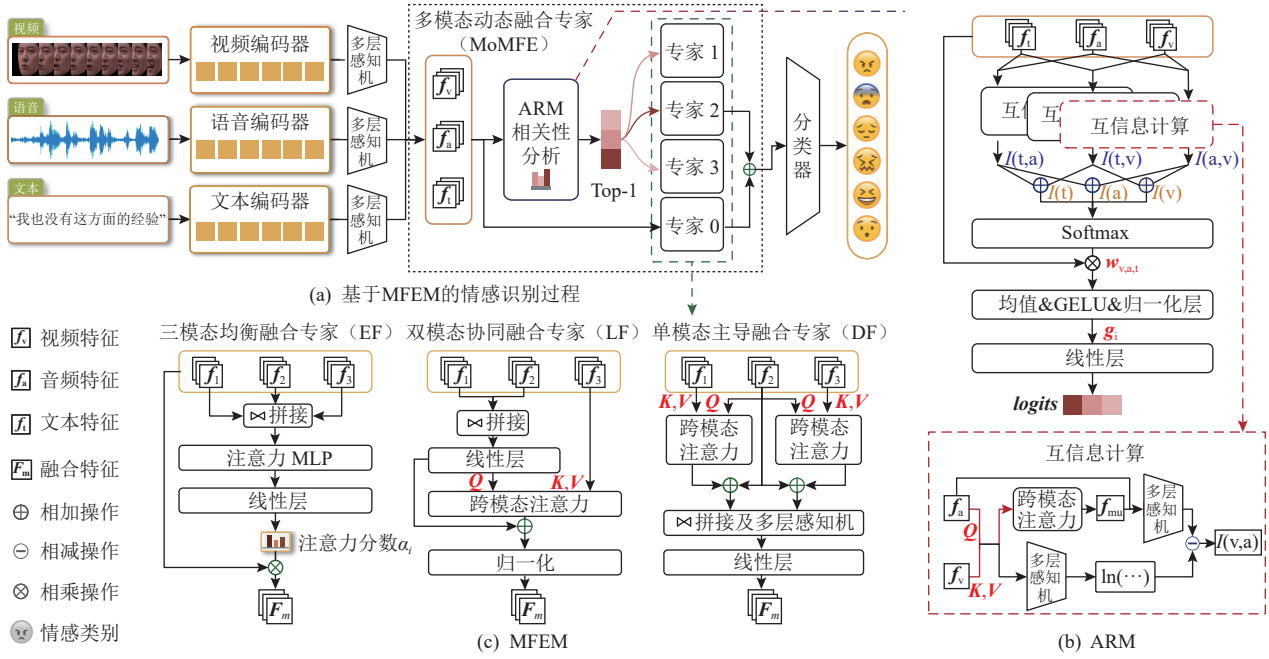


Fig. 2 The method model diagram

图 2 方法模型图

特征级别的选择方式能够更好地整合不同模态的信息,现有基于 Softmax 路由或者基于加权混合路由,根据输入多模态特征直接根据输入多模态特征直接计算样本与专家的相关性。然而面对模态贡献度不平衡的数据时,这种专家选择策略缺乏对模态间相关性的深入考量,无法保障模态贡献度组合与所选专家强相关,导致融合策略与输入特性失配。因此,如图 2(b)所示,本文设计自适应路由模块,其中模态间相关性分析部分采用互信息系数估计方法计算模态特征的相互依赖性来动态量化各模态的贡献度,然后再计算与专家的匹配分数,从而优化路由选择过程。随后,专家选择基于模态可靠性优先原则,通过建立模态可靠性评估、特征重加权以及专家激活的级联机制,确保融合过程始终聚焦于最可信的模态组合。具体计算过程如下:

1)相关性分析模块

①计算相互依赖系数。首先分别计算模态间的相互表示 $\boldsymbol{f}_{\mathrm{mu}}(\mathrm{v,a})$, $\boldsymbol{f}_{\mathrm{mu}}(\mathrm{v,t})$, $\boldsymbol{f}_{\mathrm{mu}}(\mathrm{a,t})$, 以视频模态 $\boldsymbol{f}_{\mathrm{v}}$ 和语音模态 $\boldsymbol{f}_{\mathrm{a}}$ 为例,使用跨模态的 Transformer,其中交叉注意力(cross-attention)的查询输入为 $\boldsymbol{f}_{\mathrm{v}}$,键和值输入为 $\boldsymbol{f}_{\mathrm{a}}$。跨模态相互表示如式(1)所示。

$$\boldsymbol{f}_{\mathrm{mu}}(\mathrm{v,a}) = Transformer(\boldsymbol{Q}=\boldsymbol{f}_{\mathrm{v}}, \boldsymbol{K}=\boldsymbol{V}=\boldsymbol{f}_{\mathrm{a}}) , \tag{1}$$

其中 $\boldsymbol{Q}$ 表示查询; $\boldsymbol{V}$ 表示值; $\boldsymbol{K}$ 表示键。

然后使用 MLP 分别处理跨模态和单模态表示,通过计算二者输出的差值,获得视频和语音之间的相互依赖系数 $I(\mathrm{v,a})$, 如式(2)所示。

$$I(\mathrm{v,a}) = MLP\left(\left[\boldsymbol{f}_{\mathrm{mu}}(\mathrm{v,a}) \oplus \boldsymbol{f}_{\mathrm{a}}\right]\right) - \ln\left(MLP\left(\left[\boldsymbol{f}_{\mathrm{v}} \oplus \boldsymbol{f}_{\mathrm{a}}\right]\right)\right) , \tag{2}$$

其中 $\left[\boldsymbol{f}_{\mathrm{mu}}(\mathrm{v,a}) \oplus \boldsymbol{f}_{\mathrm{a}}\right]$ 和 $\left[\boldsymbol{f}_{\mathrm{v}} \oplus \boldsymbol{f}_{\mathrm{a}}\right]$ 都表示对应位置特征相加。相应地,可以得到其他模态间的相互依赖系数 $I(\mathrm{v,t})$ 和 $I(\mathrm{t,a})$。

为了最大化每 2 个模态之间的相互依赖系数,引入损失 L_{mu}。通过最小化 L_{mu},模型能够促使不同模态之间的特征表示进一步对齐,从而增强多模态信息的一致性,损失 L_{mu} 计算如式(3)所示。

$$L_{\mathrm{mu}} = -I(\mathrm{a,v}) - I(\mathrm{v,t}) - I(\mathrm{t,a}) 。 \tag{3}$$

②计算可靠性系数。将同一模态相关的 2 个相互依赖系数相加定义为可靠性系数,以此衡量多个模态间的相关性以及对情感识别任务的表现能力。可靠性系数由式(4)计算。

$$\begin{cases} I(i) = I(i,j_1) + I(i,j_2) , \\ i,j_1,j_2 \in \{\mathrm{v,a,t}\}; i \neq j_1 \neq j_2 。 \end{cases} \tag{4}$$

通过 Softmax 将可靠性系数归一化为 $\boldsymbol{w}_i$, 将 $\boldsymbol{w}_i$ 作为权重对单模态特征进行加权,然后过归一化层重新映射为 $\boldsymbol{g}_i$, 如式(5)所示。

$$\begin{cases} \boldsymbol{w}_i = Softmax(I(i)), i \in \{\mathrm{v,a,t}\} , \\ \boldsymbol{g}_i = Mean(GELU(RMSNorm(\boldsymbol{w}_i \cdot \boldsymbol{f}_i))) , \end{cases} \tag{5}$$

其中 *RMSNorm* 为均方根归一化层, *GELU* 为高斯误差线性单元层, *Mean* 为均值处理。

2)融合专家及主导模态选择

最后将多模态特征相加为 $[\boldsymbol{g}_{\mathrm{v}} \oplus \boldsymbol{g}_{\mathrm{a}} \oplus \boldsymbol{g}_{\mathrm{t}}]$, 然后送入线性层处理,得到 ***logits*** 作为每个专家的激活率,并激活 Top-1 融合专家, 如式(6)所示。

$$\boldsymbol{logits} = Linear([\boldsymbol{g}_{\mathrm{v}} \oplus \boldsymbol{g}_{\mathrm{a}} \oplus \boldsymbol{g}_{\mathrm{t}}]) 。 \tag{6}$$

对于主导模态的选择,按照可靠性系数的计算从大到小依次将模态特征排序为 $\boldsymbol{f}_1$, $\boldsymbol{f}_2$, $\boldsymbol{f}_3$:

$$\begin{cases} \boldsymbol{p} = Sort(I(i)), i \in \{\mathrm{v,a,t}\} , \\ \boldsymbol{f}_1 = \boldsymbol{p}[0], \boldsymbol{f}_2 = \boldsymbol{p}[1], \boldsymbol{f}_3 = \boldsymbol{p}[2] 。 \end{cases} \tag{7}$$

为了进一步引导路由激活相匹配的融合专家,引入专家引导损失 L_{na}。首先将专家激活率 ***logits*** 从大到小排序为 $E=\{E_1,E_2,E_3\}$, 然后通过多模态可靠性系数 $\boldsymbol{w}_i$ 计算每个专家的置信度为 $S=\{S_1,S_2,S_3\}$, 如式(8)所示。

$$\begin{cases} r_{\mathrm{l}}, r_{\mathrm{m}}, r_{\mathrm{s}} = Sort(\boldsymbol{w}_i), i \in \{\mathrm{v,a,t}\} , \\ S_{\mathrm{EF}} = 1 - avg(r_{\mathrm{l}}, r_{\mathrm{m}}, r_{\mathrm{s}}) , \\ S_{\mathrm{LF}} = 1 - \dfrac{r_{\mathrm{s-l}}}{\max(r_{\mathrm{l-s}}, r_{\mathrm{s-l}})} , \\ S_{\mathrm{DF}} = 1 - \dfrac{r_{\mathrm{m-s}}}{\max(r_{\mathrm{l-s}}, r_{\mathrm{m-s}})} 。 \end{cases} \tag{8}$$

融合专家的 E 与 S 之间的引导损失定义为

$$L_{\mathrm{na}}(E,S) = \sum_{(i,j):\, S_i < S_j} h\left(E_j - E_i\right) , \tag{9}$$

其中函数 h 是合页损失函数 $h(x) = \max(1-x, 0)$, 该函数为非递增函数。如果 $S_j < S_i$ 则 $E_j < E_i$, 该函数惩罚 E 和 S 之间的反向配对,并鼓励 E 和 S 的顺序相同,从而引导路由选择置信度高的融合专家。

2.3 多模态融合专家层

如图 2(c)所示,本文利用跨模态注意力机制,构建了多种融合专家来处理各种多模态贡献度组合,包含单模态主导融合(dominant fusion, DF)专家、双模态协同融合(lack fusion, LF)专家、三模态均衡融合(equal fusion, EF)专家以及共享融合(shared fusion, SF)专家。其中, DF 专家适用于某一模态特征的贡献程度显著高于其他模态; LF 专家适用于某一模态特征贡献度较差的情况,此时应更加依赖其他 2 个模态进行融合; EF 专家适用于所有模态贡献度均衡的情况; SF 专家作为共享融合专家,旨在捕获独立于

模态贡献度变化的通用情感特征，为模型提供基础的泛化表示能力。

1）DF 专家

首先采用 2 个跨模态注意力模块基于主导模态进行特征增强，其中主导模态 $\boldsymbol{f}_1$ 被用作查询 $\boldsymbol{Q}$，以指定注意力机制的聚焦位置。随后，将增强后的模态特征进行拼接，并通过线性层进一步处理后与主导模态相加，最终得到融合特征 $\boldsymbol{F}_m$：

$$\begin{cases} Att(\boldsymbol{f}_1,\boldsymbol{f}_2)=Attention(\boldsymbol{Q}=\boldsymbol{f}_1,\boldsymbol{K}=\boldsymbol{f}_2,\boldsymbol{V}=\boldsymbol{f}_2)\,, \\ Att(\boldsymbol{f}_1,\boldsymbol{f}_3)=Attention(\boldsymbol{Q}=\boldsymbol{f}_1,\boldsymbol{K}=\boldsymbol{f}_3,\boldsymbol{V}=\boldsymbol{f}_3)\,, \\ \boldsymbol{F}'_m=Linear(Concat(Att(\boldsymbol{f}_1,\boldsymbol{f}_2),Att(\boldsymbol{f}_1,\boldsymbol{f}_3)))\,, \\ \boldsymbol{F}_m=\boldsymbol{f}_1+\boldsymbol{F}'_m\,。 \end{cases} \tag{10}$$

2）LF 专家

采用一个跨模态注意力模块对贡献度较差的模态 $\boldsymbol{f}_3$ 进行补充学习。其中，注意力机制的 $\boldsymbol{Q}$ 由 2 个重要模态特征的拼接 $Concat(\boldsymbol{f}_1,\boldsymbol{f}_2)$ 构成，而 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 均来自 $\boldsymbol{f}_3$。随后，通过残差连接和线性层进一步处理，得到融合特征 $\boldsymbol{F}_m$：

$$\begin{cases} \boldsymbol{F}'_m=Norm(Concat(\boldsymbol{f}_1,\boldsymbol{f}_2))\,, \\ Att(\boldsymbol{F}'_m,\boldsymbol{f}_3)=Attention\left(\boldsymbol{Q}=\boldsymbol{F}'_m,\boldsymbol{K}=\boldsymbol{f}_3,\boldsymbol{V}=\boldsymbol{f}_3\right)\,, \\ \boldsymbol{F}_m=Linear\left(\boldsymbol{F}'_m+Att\left(\boldsymbol{F}'_m,\boldsymbol{f}_3\right)\right)\,。 \end{cases} \tag{11}$$

3）EF 专家

通过 Softmax 计算模态间注意力分数进行加权融合，其计算方法如式（12）所示。其中 $\boldsymbol{F}'^{\mathrm{T}}_{\mathrm{m}}$ 表示 3 个模态特征拼接后的转置向量，$\boldsymbol{\alpha}_i$ 为注意力分数，用于衡量不同模态特征的重要性权重。$\boldsymbol{W}_i$，$\boldsymbol{b}_i$ 分别为可训练的注意力矩阵和偏置项。最终，$\boldsymbol{F}_{\mathrm{m}}$ 是融合 3 个较高质量模态特征得到的融合特征：

$$\begin{cases} \boldsymbol{F}'_m=Concat(\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3)\,, \\ \boldsymbol{\alpha}_i=Softmax\left(\boldsymbol{F}'^{\mathrm{T}}_m\cdot\boldsymbol{W}_i+\boldsymbol{b}_i\right)\,, \\ \boldsymbol{F}_m=\boldsymbol{F}'_m\cdot\boldsymbol{\alpha}_i\,。 \end{cases} \tag{12}$$

4）SF 专家

虽然在前期训练可选择的专家时，强调各个专家的独特性，即它们所学习的知识侧重不同，但是由于专家之间没有共享机制，导致专家参数的冗余。受 Deepseek-MoE[35] 启发，引入共享的融合专家，不论路由选择哪个融合专家，多模态特征都将同时分配给共享的融合专家，融合结构与均衡模态融合专家一致。最终的融合结果如式（13）所示，其中 α 为共享融合专家的重要性权重。

$$\boldsymbol{F}_{\mathrm{fusion}}=\boldsymbol{F}_m+\alpha\cdot SharedFusion(\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3)\,。 \tag{13}$$

2.4 情感识别

获得融合特征之后，将其输入情感分类器以得到情感识别的类别结果，如式（14）所示。

$$\hat{\boldsymbol{y}}=Classifier(\boldsymbol{F}_{\mathrm{fusion}})\,。 \tag{14}$$

并使用交叉熵损失作为情感分类任务的损失：

$$L_{\mathrm{task}}=-N\ln(\boldsymbol{logits}_{\mathrm{e}})\,, \tag{15}$$

其中 N 表示真实情感标签的索引，$\boldsymbol{logits}_{\mathrm{e}}$ 表示对所有情感类别预测的概率分布。

本文最终的损失函数计算如式（16）所示。

$$L=L_{\mathrm{task}}+\beta L_{\mathrm{mu}}+\gamma L_{\mathrm{na}}\,, \tag{16}$$

其中 β 和 γ 表示为总损失中各部分贡献的权重系数。

3 实验设置

3.1 数据集

本文在 3 个公开的多模态情感数据集上进行实验，都包含视频、语音、文本 3 个模态。分别是 MER2024[5]，CMU-MOSEI[36]，CH-SIMS[37]，数据集的统计细节如表 1 所示。

Table 1 Introduction of Datasets

表 1 数据集介绍

数据集	训练集	验证集	测试集	语言
MER2024	5 030		1 169	中文
CMU-MOSEI	16 326	1 871	4 659	英文
CH-SIMS	1 368	456	457	中文

MER2024 是第 2 届多模态情感识别挑战赛推出的数据集，重点关注开放词汇情感识别和半监督学习。MER2024-SEMI 数据集包含约 6 199 个有标签的视频片段，采用 6 类离散情感标签。数据集设计强调真实场景的复杂性，多来源于影视片段。

CMU-MOSEI 是多模态情感分析领域的经典数据集，包含 1 000 名不同讲述者的 22 856 个带标注的 YouTube 视频评论片段，覆盖 250 个主题，总时长约 65 h。每个片段均标注 −3 ~ +3的连续情感强度值，以及 7 种离散情绪类别。

CH-SIMS 是一个中文多模态情感数据集，专注于提供细粒度的模态注释，支持单模态和多模态情感分析任务。包含 2 281 个视频片段，采集自电影、电视剧和综艺节目，具有丰富的人物背景、年龄跨度大、数据质量高等特点。与前 2 个数据集不同的是，对于每个视频片段不仅进行多模态标注，还有单模态标注。

3.2 评价准则与评价指标

1)评价准则

本文以加权 $F1$ 分数作为核心评估指标,应对样本类别不平衡的数据集。实验过程中,所有数据集均以 $F1$ 分数作为模型选择的主要依据。为保持与其他研究的可比性,同时提供了准确率(Acc)、平均绝对误差(Mae)和相似度系数(Corr)等多项评估指标。对于 MER2024 数据集,由于官方未指定验证集,遵循官方的评价标准,对训练集采用 5 折交叉验证,并使用测试集进行模型测试。评价指标有加权平均 $F1$ 分数和 6 分类准确率(Acc-6)。对于 CMU-MOSEI 数据集和 CH-SIMS 数据集,评价指标采用 2 级精度准确率(Acc-2), Mae, Corr, $F1$。不同的是, CH-SIMS 还使用 3 级精度准确率(Acc-3)和 5 级精度准确率(Acc-5),而 CMU-MOSEI 还使用 7 级精度准确率(Acc-7)。除 Mae 外,其他指标的数值越高表明模型性能越好。

2)评价指标

Mae 是一种衡量预测值与真实值之间绝对差异的回归指标,数值越小表示模型预测越精准,在多模态情感分析中常用于评估连续情感维度(如情感强度)的预测准确性。

Corr 用于评估预测值与真实值之间的线性相关性,其值越接近 1 表示模型对情感趋势的捕捉能力越强。能有效反映模型是否保留了情感变化的相对顺序,但对绝对误差不敏感,常与 Mae 互补使用。

Acc 是分类任务中最直观的评估指标,表示正确预测的样本比例,不同细粒度的准确率可反映模型从基础极性判别到细粒度分类的综合能力,但在数据类别不平衡时可能会掩盖模型对少数类的识别缺陷。

$F1$ 综合了精确率与召回率的调和平均值,适合处理类别不平衡的情感数据集,如某些情绪样本稀少的情况。$F1$ 通过按类别样本量加权,能更公平地评估模型对各类情感的识别均衡性,是多模态情感任务的核心指标,其提升通常直接反映模型实用性的增强。

3.3 实验参数设置

实验环境基于 Python3.9 和 PyTorch2.3 构建,所有实验均在配备 NVIDIA GeForce RTX 3 090 GPU 的设备上进行。模型训练采用 Adam 优化器,其权重衰减为 0.02,初始学习率为 0.000 1,最大迭代次数为 40,批大小设置为 4。分类任务使用交叉熵损失函数,回归任务使用均方误差损失函数。其他超参数设置:各模态特征经 MLP 映射到特征维度 $d = 256$,共享专家权重为 $\alpha = 0.15$,损失 L_{mu} 权重为 $\beta = 0.2$,损失 L_{na} 权重为 $\gamma = 0.2$。

3.4 特征提取

对于视觉特征,先使用 openFace 工具箱[38]从视频中检测并对齐人脸,再通过 CLIP-large 图像编码器[39]提取深度特征。对于声学特征,与之前的工作[40]相同,使用 HuBert-large 模型[18]提取深度特征。对于文本特征,使用 Baichuan2-13B 模型[41]提取深度特征。视觉特征、音频特征和文本特征的维度分别为 $d_v = 768$, $d_a = 1\,024$, $d_t = 5\,120$。

4 实验结果与分析

本文设计了以下 6 方面的实验以全面验证所提方法的有效性:与其他前沿方法的对比实验、自适应路由有效性实验、多模态融合专家有效性实验、缺失模态下模型性能实验、模型噪声鲁棒性实验、超参数评估实验。

4.1 与其他前沿方法的对比实验

为全面评估本文方法的有效性,本文在 MER2024, CMU-MOSEI, CH-SIMS 这 3 个情感数据集上,与当前最先进的 12 种多模态融合方法进行了系统对比。对比方法涵盖 3 个方面:1)基于固定模态主导的方法(MPID[42], GFNPME[43], MR-CCL[44]);2)动态权重融合方法(SIMSUF[4], DWE[7], KuDA[8], LAA[40], AGT[45]);3)混合专家架构方法(GLoMo[10], TGMoE[11], FuseMoE[32], AC-MoE[46]),实验数据来源于所对比文章的公开结果。在 MER2024 数据集的实验结果如表 2 所示,在 CMU-MOSEI 和 CH-SIMS 数据集的实验结果如表 3 所示。本文方法 MoMFE 在情感识别准确率(Acc-2)、$F1$ 分数等核心指标上较所有对比方法均有提升。

Table 2 Results of Different Methods on MER2024 Dataset

表 2 不同方法在 MER2024 数据集上的结果 %

方法	训练及验证 $F1$	测试 $F1$	Acc-6
MR-CCL	**86.53**	88.25	
LAA	83.85	88.90	
AGT	83.22	89.83	
MoMFE(本文)	83.28	**89.88**	**90.01**

注:黑体数值为最优值。

1)与基于固定模态主导方法对比

与 MR-CCL, MPID, GFNPME 方法进行对比。其中, MR-CCL 通过固定权重结合多模态和单模态分类

Table 3 Results of Different Methods on CMU-MOSEI and CH-SIMS Datasets

表 3 不同方法在 CMU-MOSEI 和 CH-SIMS 数据集上的结果 %

方法	CMU-MOSEI 数据集					CH-SIMS 数据集					
	Mae	Corr	Acc-7	Acc-2	*F*1	Mae	Corr	Acc-5	Acc-3	Acc-2	*F*1
MPID	53.3	77.6	53.64	85.10	85.30	42.1	61.8	**44.21**		81.56	81.91
GFNPME	52.8	77.1	53.90	86.52	86.26	40.2	61.6			83.15	83.10
SIMSUF	52.9	77.2	53.68	86.23	86.12						
KuDA	52.9	77.6	52.89	86.46	86.59	40.8	61.3	43.54	66.52	80.74	80.71
DWE	**52.6**	76.8	54.80	85.97	85.98	41.6	59.3	42.90		80.74	80.54
AC-MoE						**31.5**	73.2			82.05	82.04
TGMoE	53.5	75.7	53.70	85.51	85.86						
FuseMoE	54.0	**81.0**		86.41	85.43						
GLoMo	53.9	77.1	**55.00**	86.50	86.40						
MoMFE（本文）	53.4	78.8	54.72	**86.83**	**86.69**	37.7	**75.1**	41.38	**71.99**	**84.33**	**84.51**

注：黑体数值为最优值。

器的输出；MPID 首先以文本模态为主导分别整合音频和视觉模态，最后进行全局融合；GFNPME 先利用其他模态增强文本表征，再逐步优化视觉和音频模态的特征。实验数据表明，在 MER2024 数据集中，与采用固定权重融合的 MR-CCL 方法相比，MoMFE 在 *F*1 分数上提升了 1.63 个百分点。在 CMU-MOSEI 和 CH-SIMS 数据集中，MoMFE 在 Corr, Acc-2, *F*1 分数等评价指标上均优于以文本为主导模态的固定融合方法（MPID 和 GFNPME）。这些结果说明本文提出的 ARM 能够有效地动态衡量各模态的贡献度，在多模态融合中实现更全面的评估。

2）与动态权重融合方法对比

与 LAA, AGT, SIMSUF, KuDA, DWE 等方法进行比较。LAA 计算模态间注意力分数进行加权融合，AGT 基于多头注意力机制计算模态间相似性，并在某一模态与其他模态相似性显著偏低时直接剔除该模态；SIMSUF, KuDA, DWE 方法分别通过计算互信息系数、动态注意力融合模块和遗传算法，动态选择主导模态进行融合。从实验数据综合来看，MoMFE 在 3 个数据集上多个指标均优于以上方法，其中在 CH-SIMS 数据集上提升更为明显，与单一的融合策略的 KuDA 方法相比，Acc-2 和 *F*1 指标分别提升了 3.59% 和 3.80%，这一结果说明了多种互补融合方法的设计为数据提供更灵活的选择。在 MER2024 数据集中，相较于 AGT 方法直接剔除贡献度低的模态，MoMFE 方法通过特征补充使得 *F*1 分数提升了 0.05%。这表明，MoMFE 能够进一步挖掘低贡献度模态的潜在信息。

3）与混合专家架构方法对比

与 AC-MoE, TGMoE, FuseMoE, GLoMo 方法进行比较。这些方法都使用 MoE 层实现多模态融合，但各有侧重。AC-MoE 中的 MoE 网络用于提取融合表示中非语言模态的信息；TGMoE 中的 MoE 层用来增强以文本主导的融合模态中的情感信息；FuseMoE 设计了特殊的路由来控制专家的细粒度的多模态数据输入；GLoMo 中的 MoE 层仅用于单模态的局部表示模块。与以上几种方法相比，MoMFE 在 CMU-MOSEI 数据集上 Acc-2 和 *F*1 指标都提高了 0.3% 以上，在 CH-SIMS 数据集上 Acc-2 和 *F*1 指标比 AC-MoE 方法提升 2.28% 和 2.47%，充分证明了将 MoE 用于设计动态选择多模态融合方法的有效性。

综合来看，MoMFE 在 Acc-2 和 *F*1 指标上取得了最优效果，同时在部分指标上与现有最优方法保持接近。具体而言，在 CMU-MOSEI 数据集上，其 Corr 指标低于 FuseMoE, Acc-7 指标与 GLoMo 的差距仅为 0.28%；在 CH-SIMS 数据集上，虽然 Mae 值优于大多数对比方法，但与 AC-MoE 相比仍有差距。这些实验数据表明，MoMFE 在连续情感识别指标和细粒度分类指标上仍有提升空间，但在 *F*1 分数这一关键指标上展现了显著优势，这一结果充分说明了 MoMFE 在处理类别不平衡数据集时的优越性能。值得注意的是，MoMFE 在 CH-SIMS 数据集上的性能提升显著优于其他基准数据集（MER2024, CMU-MOSEI）。这一差异可能源于各数据集内的模态特性分布：CH-SIMS 数据集场景复杂，呈现出更加均衡的跨模态贡献度分布；相比之下，MER2024 数据集中

的情感表达主要依赖于声学模态特征[5], 而 CMU-MOSEI 数据集则存在明显的文本模态主导倾向[37,47]。MoMFE 在 CH-SIMS 数据集上的结果说明其针对真实场景中多变的多模态融合挑战所具备的适应性和性能优势。

4.2 自适应路由机制的有效性实验

自适应路由中的关键组成部分是相关性分析模块, 因此, 为验证该模块的有效性, 设计如下实验:

1)在 MER2024 和 CH-SIMS 数据集上, 对比了 3 种路由设置的效果: 传统的 Softmax 路由、本文提出的自适应路由机制(ARM), 以及在 ARM 基础上引入专家引导损失 L_{na}。其中 Softmax 路由直接计算专家激活率 ***logits***, 并不包含相关性分析模块, 实验结果如表 4 所示。

Table 4 Experimental Results Under Different Router Settings

表 4 不同路由设置下的实验结果 %

路由机制	MER2024 数据集		CH-SIMS 数据集	
	Acc-6	*F*1	Acc-2	*F*1
Softmax 路由	87.68	87.45	81.14	81.19
ARM 路由	89.39	88.81	83.36	83.11
ARM 路由+L_{na}	**90.01**	**89.88**	**84.33**	**84.51**

注: 黑体数值为最优值。

表 4 系统地评估了 3 种路由机制在跨模态情感识别任务中的表现。首先, 相较于传统 Softmax 路由(直接计算专家激活率而未深入考虑模态相关性), 本文提出的 ARM 路由通过引入模态相关性分析模块, 在 2 个数据集上都取得了性能提升: MER2024 数据集的 *F*1 分数从 87.45% 提升至 88.81%, CH-SIMS 数据集从 81.19% 提升至 83.11%。这一性能提升得益于 ARM 路由引入的相关性分析模块, 能够基于模态间的深层交互实现更有效的专家选择。相比之下, Softmax 路由仅基于浅层的模态理解直接进行专家选择, 在处理模态间复杂交互时存在一定的局限性。其次, 进一步引入 L_{na} 损失函数后, 模型性能获得持续提升, 即 MER2024 和 CH-SIMS 数据集的 *F*1 分数分别达到 89.88% 和 84.51%。这一结果说明, L_{na} 通过约束路由决策与专家能力的一致性, 强化了“模态可靠性-专家专业化”的匹配关系, 促使路由形成更清晰的决策边界。

2)为了更直观地观察路由的分配数据的能力, 可视化不同路由设置下每个融合专家所处理的数据类型比率(单模态主导数据、双模态协同数据、三模态均衡数据), 如图 3 所示。

从图 3 可以看出, 采用 Softmax 路由时, 数据分布没有与特定专家的能力相匹配。例如, 三模态均衡类型样本被分散到 3 个专家, 而单模态主导类型样本却更多被分配给了 EF 专家中, 并未依据模态贡献度组合将样本输入到相应的融合专家中。然而, 在使用引入了相关性分析模块的 ARM 路由后, 可以明显观察到路由对模态间相关性的感知能力增强。例如, DF 专家更多被单模态主导类型样本激活, LF 专家更多被双模态协同类型样本激活, EF 专家更多被三模态均衡类型样本激活。此外, 这一数据也体现了在添加相关性模块之后, 路由对专家选择的正确率有明显提升。整体来看, 这一分配模式与各专家的设计初衷高度一致, 证明了相关性分析模块能够增强路由对模态贡献模式的判别能力。

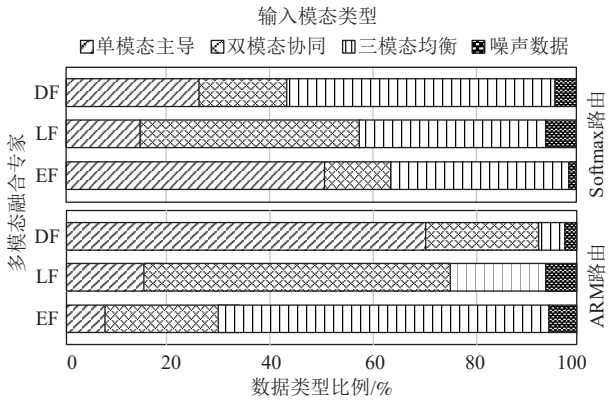


Fig. 3 Expert activation rates under different router settings on CH-SIMS dataset

图 3 CH-SIMS 数据集上不同路由设置下的专家激活率

此外, 提出的基于相关性分析模块的自适应路由机制, 不仅是 ARM 路由性能提升的核心原因, 其本身也仅引入极少的参数与计算开销。该模块通过轻量化的模态交互计算, 显著提升了路由决策精度, 为后续专家的高效调度奠定了坚实基础。

4.3 多模态融合专家的有效性实验

为验证多模态融合专家的性能, 设计如下实验:

1)为评估本文构建的 3 个互补融合专家在处理不同模态贡献组合时的性能, 将 CH-SIMS 数据集分为 3 类数据分别是: 单模态主导类型、双模态协同类型、三模态均衡类型, 各自输入到专家中进行特征融合, 然后进行情感识别并计算 *F*1 分数。实验结果如图 4 所示。

如图 4 所示, 三模态均衡类型样本在所有专家中的 *F*1 分数均高于其他类型样本, 这表明模态贡献度的不平衡确实对情感识别任务产生了影响。具体

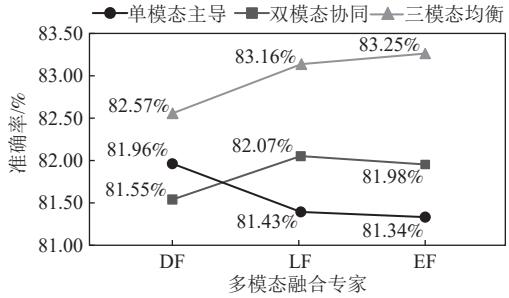


Fig. 4 Performance of each fusion expert on CH-SIMS classification dataset

图 4 CH-SIMS 分类数据集上各融合专家的表现

来看，每种数据类型在最适合其特点的融合方法下性能最佳。例如，EF 专家在三模态均衡数据上获得 83.25% 的 *F*1 分数，显著高于其在其他数据类型的表现。这归因于其采用的多模态交叉注意力机制能够平等地整合所有模态信息，避免因强调单一模态而损失互补信息。LF 专家在双模态协同数据上表现最佳，*F*1 分数为 82.07%，这是由于其设计的双模态交互的注意力机制能够识别并强化 2 个主导模态间的协同效应。DF 专家在单模态主导数据上取得 81.96% 的 *F*1 分数，这得益于其主导模态增强的注意力机制，突出高贡献模态的特征的同时抑制噪声模态的干扰。这些结果证明，本文设计多样化互补融合方法能够最大化利用贡献度不均衡的数据，提升情感识别的准确性。

2）为评估 SF 专家的性能，在 CH-SIMS 数据集上进行消融实验，包含 4 种融合专家。首先在完整专家组合（DF+LF+EF）基础上进行 SF 专家的消融对照，其次针对专家两两组合（LF+EF, DF+EF, DF+LF）分别进行 SF 专家的消融测试，进一步全面考察 SF 专家在不同专家组合条件下的性能表现。实验结果如表 5 所示。

如表 5 所示，实验结果展现了不同专家组合在 CH-SIMS 数据集上的性能差异。通过系统性地对比分析 DF, LF, EF, SF 这 4 种融合专家的组合效果，发现模型性能随着专家组合的完善而显著提升，其中在 DF+LF+EF 基础上添加 SF 专家后 *F*1 分数提升 0.8%，并在 3 项指标上均达到最优，说明 SF 专家通过参数共享机制有效学习了跨专家的通用融合模式，从而增强了对常见情感表达规律的捕获能力，并提升了模型的泛化能力。值得注意的是，即便在部分专家缺失的情况下，引入 SF 的专家组合仍能维持相对稳定的表现，这表明 SF 专家可作为“安全专家”提供稳健的输出，有效缓解因专家缺失导致的性能下降，确保系统稳定性。

Table 5 Experimental Results of Different Expert Combinations Setting

表 5 不同专家组合设置的实验结果 %

多模态融合专家				CH-SIMS 数据集		
DF	LF	EF	SF	Acc-3	Acc-2	*F*1
√	√			68.19	80.39	80.37
√	√		√	69.37	81.33	81.53
√		√		69.10	82.35	82.39
√		√	√	70.48	82.39	82.77
	√	√		70.07	82.13	82.55
	√	√	√	70.39	82.53	82.68
√	√	√		71.25	83.42	83.71
√	√	√	√	**71.99**	**84.33**	**84.51**

注：黑体数值为最优值。

此外，得益于前置路由的高效分配，对于每个输入样本，模型仅激活最匹配的融合专家，而无需同时计算所有专家网络。这种专家激活机制，使得模型在绝大部分情况下实际参与计算的参数量被大幅缩减，从而在保证模型性能提升的同时，有效控制计算开销。

4.4 缺失模态下模型性能实验

为测试 MoMFE 在不同缺失模态情况下的鲁棒性表现，进行如下实验：模拟现实场景中常见的模态不完整情况，在 MER2024, CMU-MOSEI, CH-SIMS 这 3 个数据集上分别设计单模态缺失和双模态缺失这 2 种场景，设置了 10%, 30%, 50% 这 3 种不同比例的随机缺失样本。该实验主要观察模型在 Acc 和 *F*1 这 2 个核心指标的性能变化。为确保实验结果的可靠性，所有实验条件均重复 5 次并取均值作为最终结果，实验结果如表 6 所示。

从表 6 看出，不同数据集对模态缺失的敏感性存在显著差异。MER2024 和 CMU-MOSEI 数据集受模态缺失的影响更为明显，在单模态 50% 缺失情况下，MoMFE 的 *F*1 分数在 MER2024 数据集上相比基线下降了 3.75%，相比 CMU-MOSEI 数据集下降了 2.32%。而相比 CH-SIMS 数据集仅下降 1.67%。这可能是由于 MER2024 和 CMU-MOSEI 数据集对单模态的依赖性更强，当这些关键模态缺失时，模型难以从剩余模态中获得足够的互补信息。且双模态缺失的情况会进一步加剧指标下降，这在 3 个数据集的实验结果中都得到了验证。同时，即便在严重的模态缺失条件下，MoMFE 方法仍能保持相对稳定的性能表现，

Table 6 Experimental Results of Model Performance with Missing Modalities

表 6 缺失模态下模型性能的实验结果 %

模态缺失类型	缺失样本比例	MER2024 数据集		CMU-MOSEI 数据集		CH-SIMS 数据集	
		Acc	*F*1	Acc-2	*F*1	Acc-2	*F*1
单模态缺失	10	89.27±0.07	88.83±0.12	86.12±0.12	86.25±0.08	83.81±0.11	84.03±0.14
	30	88.01±0.12	87.55±0.17	84.89±0.24	85.02±0.27	83.11±0.17	83.37±0.21
	50	86.43±0.10	86.13±0.05	84.14±0.20	84.37±0.16	82.69±0.26	82.84±0.19
双模态缺失	10	87.32±0.03	87.10±0.10	84.26±0.13	84.53±0.12	83.28±0.12	83.81±0.15
	30	86.52±0.11	86.31±0.18	83.52±0.24	83.88±0.17	82.03±0.06	82.44±0.09
	50	84.35±0.09	83.87±0.23	83.04±0.15	83.29±0.18	81.42±0.25	81.75±0.14

注: “±”前数值表示平均值, “±”后数值表示误差。

特别是在 CH-SIMS 数据集上展现出较强的鲁棒性, 这充分证明了 MoMFE 在非理想数据条件下的适应能力和实用价值。

4.5 模型噪声鲁棒性测试

为测试 MoMFE 在噪声环境下的鲁棒性, 进行如下实验: 在 3 个模态上采用与 MER2024 官方一致的加噪方式, 对于视觉模态采用高斯模糊模拟图像退化, 对于语音模态通过添加环境噪声再现真实场景中的语音干扰, 而文本模态保留 ASR 自动识别引入的固有噪声数据。为模拟现实场景中的噪声情况, 实验在 MER2024 数据集上分别设计单模态加噪以及多模态加噪, 并设置了 10%, 30%, 50% 共 3 种不同的随机加噪的样本比例。为确保实验结果的可靠性, 所有实验条件均重复 5 次并取均值作为最终结果, 实验结果如表 7 所示。

Table 7 Experimental Results of Noise Robustness

表 7 噪声鲁棒性实验结果 %

模态加噪类型	加噪样本比例	MER2024 数据集	
		Acc	*F*1
语音模态加噪	10	88.57±0.15	88.21±0.13
	30	87.44±0.05	87.10±0.08
	50	86.23±0.14	85.95±0.15
视觉模态加噪	10	89.11±0.08	88.85±0.10
	30	88.64±0.06	88.26±0.05
	50	87.34±0.11	87.05±0.13
语音+视觉加噪	10	88.23±0.07	87.94±0.09
	30	87.29±0.15	87.12±0.11
	50	85.86±0.08	85.62±0.05

注: “±”前数值表示平均值, “±”后数值表示误差。

表 7 所示的 MER2024 数据集实验结果表明, MoMFE 在不同噪声条件下均表现出良好的鲁棒性。语音模态作为该数据集的重要模态, 其加噪 10%~50% 时模型性能指标 *F*1 分数从 88.21% 降至 85.95%, 下降幅度较视觉模态加噪情况更为明显。值得注意的是, 即使在语音+视觉双模态同时加噪且样本比例为 50% 的情况下, 模型指标 *F*1 分数仍能保持 85.62% 的稳定表现, 仅比单语音模态加噪时下降约 0.4 个百分点。这一结果表明, 虽然语音模态的噪声对模型性能影响较大, 但整体而言, 模型在不同类型和程度的噪声干扰下都表现出较为稳定的适应能力。

4.6 超参数评估实验

为了研究特征维度 d 以及共享专家权重 α 对 MoMFE 模型性能的影响, 本文设置特征维度 d 取值为 128, 256, 512, α 范围为 0.1~0.3。图 5 展示了不同设置下 MoMFE 模型的 *F*1 分数。

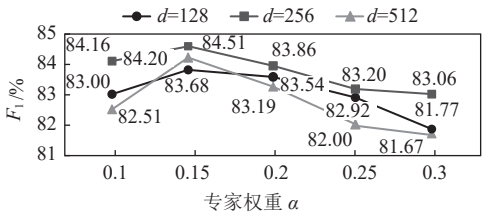


Fig. 5 Assessment results of d and α for MoMFE on CH-SIMS dataset

图 5 CH-SIMS 数据集上 d 与 α 对 MoMFE 的评估结果

图 5 的实验数据展示了特征维度 d 和共享专家权重 α 对模型性能的影响。当特征维度 d=256 时, 模型整体表现最优, 最高 *F*1 分数达 84.51%, 表明该特征维度在表征能力和计算效率之间较为平衡: 较小的特征维度(d=128)导致特征表达能力不足, 最高 *F*1 分数仅 83.68%, 而过大的特征维度(d=512)则可能引入噪声, 性能波动明显, 最高 84.20% 降至最低 81.67%。对于共享专家权重 α, 实验显示在 0.1~0.2 表现效果显著, 其中 α=0.15 时在 d=256 条件下取得峰值; 当 α 超过 0.2 后, 性能呈单调下降趋势, 说明过大

的专家共享权重会影响各融合专家的特异性，而过分依赖共享融合专家的输出会影响情感识别的准确性。

5 结　论

本文提出一种基于混合专家框架的多模态情感识别方法，旨在解决多模态数据中情感特征质量不平衡的核心问题，提升情感识别任务的准确性与结果一致性。该方法的核心在于设计了多模态融合专家层，该层级由自适应路由模块与多模态重要性融合专家模块构成。一方面，自适应路由模块通过引入模态间相关性分析度量，实现对各模态动态贡献度的量化，同时结合专家引导损失函数优化路由决策机制，从而高效完成对最优融合专家的自适应选择；另一方面，针对不同模态在情感表征中的贡献差异，多模态重要性融合专家模块通过强化模态间互补信息的挖掘与利用，有效缓解了多模态数据中情感特征质量不均衡的问题，进一步提升了模型对复杂情感场景的表征能力。实验结果表明，相较于现有方法，本文所提基于混合专家的多模态情感识别方法在多模态情感识别任务中展现出显著的性能优势，验证了该方法在提升情感识别准确性与结果稳定性方面的有效性，为多模态情感分析领域提供了具有参考价值的技术思路。

作者贡献声明：陈海丰提出方法思路以及修改意见；李溪实现方法设计并撰写论文；李岩负责部分实验以及论文修改；李健负责实验监督以及论文修改；何浪负责项目管理以及论文修改；蒋冬梅负责项目监督以及论文修改。

参　考　文　献

[1] Das R, Singh T D. Multimodal sentiment analysis: A survey of methods, trends, and challenges[J]. ACM Computing Surveys, 2023, 55(13): 1−38

[2] Poria S, Hazarika D, Majumder N, et al. Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research[J]. IEEE Transactions on Affective Computing, 2020, 14(1): 108−132

[3] Islam M S, Kabir M N, Ghani N A, et al. Challenges and future in deep learning for sentiment analysis: A comprehensive review and a proposed novel hybrid approach[J]. Artificial Intelligence Review, 2024, 57(62): 1−79

[4] Huang Jian, Ji Yanli, Qin Zhen, et al. Dominant single-modal supplementary fusion (SIMSUF) for multimodal sentiment analysis[J]. IEEE Transactions on Multimedia, 2023, 26: 8383−8394

[5] Lian Zhen, Sun Haiyang, Sun Licai, et al. MER2024: Semi-supervised learning, noise robustness, and open-vocabulary multimodal emotion recognition[C]//Proc of the 2nd Int Workshop on Multimodal and Responsible Affective Computing. New York: ACM, 2024: 41−48

[6] Luo Yuanyi, Wu Rui, Liu Jiafeng, et al. Multimodal sentiment analysis method for sentimental semantic inconsistency[J]. Journal of Computer Research and Development, 2025, 62(2): 374−382 (in Chinese)
(罗渊贻, 吴锐, 刘家锋, 等. 面向情感语义不一致的多模态情感分析方法[J]. 计算机研究与发展, 2025, 62(2): 374−382)

[7] Liang Yang. A dynamic weighted fusion model for multimodal sentiment analysis[J]. Signal, Image and Video Processing, 2025, 19(8): 609−623

[8] Feng Xinyu, Lin Yuming, He Lihua, et al. Knowledge-Guided dynamic modality attention fusion framework for multimodal sentiment analysis[J]. arXiv preprint, arXiv: 2410.04491, 2024

[9] Shen Jintao. Research on multimodal emotion recognition based on modal interaction and dynamic fusion[D]. Shenyang: Liaoning University, 2024 (in Chinese)
(申锦涛. 基于模态交互和动态融合的多模态情感识别研究[D]. 沈阳: 辽宁大学, 2024)

[10] Zhuang Yan, Zhang Yanru, Hu Zheng, et al. GLoMo: Global-local modal fusion for multimodal sentiment analysis[C]//Proc of the 32nd ACM Int Conf on Multimedia. New York: ACM, 2024: 1800−1809

[11] Zhao Xueliang, Wang Mingyang, Tan Yingchun, et al. TGMoE: A text guided mixture-of-experts model for multimodal sentiment analysis[J]. International Journal of Advanced Computer Science & Applications, 2024, 15(8): 1227−1234

[12] Jia Xibin, Zeng Meng, Mi Qing, et al. Domain alignment adversarial unsupervised cross-domain text sentiment analysis algorithm[J]. Journal of Computer Research and Development, 2022, 59(6): 1225−1270 (in Chinese)
(贾熹滨, 曾檬, 米庆, 等. 领域对齐对抗的无监督跨领域文本情感分析算法[J]. 计算机研究与发展, 2022, 59(6): 1225−1270)

[13] Chen Haifeng, Zhang Peng, Guo Chujia, et al. Facial action unit representation based on self-supervised learning with ensembled priori constraints[J]. IEEE Transactions on Image Processing, 2024, 33: 5045−5069

[14] Wang Yuhua, Shen Guang, Xu Yuezhu, et al. Learning mutual correlation in multimodal transformer for speech emotion recognition[C/OL]//Proc of the 2021 Interspeech. 2021: 4518−4522.[2025-04-16]. https://www.isca-archive.org/interspeech_2021/wang21ga_interspeech.html

[15] Pennington J, Socher R, Manning C, et al. GloVe: Global vectors for word representation[C]//Proc of the 2014 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2014: 1532−1543

[16] Devlin J, Chang Mingwei, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proc of the 2019 Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2019: 4171−4186

[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint, arXiv: 2010.11929, 2020

[18] Hsu W N, Bolte B, Tsai Y H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451−3460

[19] Radford A, Kim J W, Xu Tao, et al. Robust speech recognition via large-scale weak supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2023: 28492−28518

[20] Wei Han, Chen Hui, Poria S. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis[C]//Proc of the 2021 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2021: 9180–9192

[21] Cai Yunrui, Wu Zhiyong, Jia Jia, et al. Lora-Mer: Low-rank adaptation of pre-trained speech models for multimodal emotion recognition using mutual information[C/OL]//Proc of the 2024 Interspeech. 2024: 4658−4662.[2025-04-15]. https://www.isca-archive.org/interspeech_2024/cai24b_interspeech.html

[22] Kumar A, Vepa J. Gated mechanism for attention based multi modal sentiment analysis[C]//Proc of the 2020 IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2020: 4477−4481

[23] Majumder N, Hazarika D, Gelbukh A, et al. Multimodal sentiment analysis using hierarchical fusion with context modeling[J]. Knowledge-Based Systems, 2018, 161: 124−133

[24] Jacobs R A, Jordan M I, Nowlan S J, et al. Adaptive mixtures of local experts[J]. Neural Computation, 1991, 3(1): 79−87

[25] Fedus W, Zoph B, Shazeer N. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity[J]. Journal of Machine Learning Research, 2022, 23(120): 1−39

[26] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proc of the 31st Int Conf on Neural Information Processing Systems. New York: ACM, 2017: 6000−6010

[27] Lepikhin D, Lee H J, Xu Yuanzhong, et al. Gshard: Scaling giant models with conditional computation and automatic sharding[C/OL]// Proc of the 9th Int Conf on Learning Representations. 2021: 1−23. [2025-04-23]. https://openreview.net/forum?id=qrwe7XHTmYb

[28] Du Nan, Huang Yanping, Dai A M, et al. GlaM: Efficient scaling of language models with mixture-of-experts[C]//Proc of the 39th Int Conf on Machine Learning. New York: ACM, 2022: 5547−5569

[29] Cai Weilin, Jiang Juyong, Wang Fan, et al. A survey on mixture of experts in large language models[J]. IEEE Transactions on Knowledge and Data Engineering, 2025, 37(7): 3896−3915

[30] Sarkar R, Liang Hanxue, Fan Zhiwen, et al. Edge-MoE: Memory-efficient multi-task vision transformer architecture with task-level sparsity via mixture-of-experts[C]//Proc of the 2023 IEEE/ACM Int Conf on Computer Aided Design. Piscataway, NJ: IEEE, 2023: 1−9

[31] Lin X V, Shrivastava A, Liang Luo, et al. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts[J]. arXiv preprint, arXiv: 2407.21770, 2024

[32] Han Xing, Nguyen H, Harris C, et al. FuseMoE: Mixture-of-experts transformers for fleximodal fusion[C]//Proc of the 38th Int Conf on Neural Information Processing Systems, New York: ACM, 2025: 67850–67900

[33] Xiong Congxiong, Chen Hao, Zheng Hao, et al. Mome: Mixture of multimodal experts for cancer survival prediction[C]//Proc of the 27th Int Conf on Medical Image Computing and Computer-Assisted Intervention. Springer: Berlin, 2024: 318−328

[34] Shazeer N, Mirhoseini A, Maziarz K, et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer[J]. arXiv preprint, arXiv: 1701.06538, 2017

[35] Dai Damai, Deng Chengqi, Zhao Chenggang, et al. Deepseek-Moe: Towards ultimate expert specialization in mixture-of-experts language models[C]//Proc of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2024: 1290−1297

[36] Zadeh A B, Liang P P, Poria S, et al. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2018: 2236−2246

[37] Yu Wenmeng, Xu Hua, Meng Fanyang, et al. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 3718−3727

[38] Baltrusaitis T, Zadeh A, Lim Y C, et al. OpenFace 2.0: Facial behavior analysis toolkit[C]//Proc of the 13th IEEE Int Conf on Automatic Face & Gesture Recognition. Piscataway, NJ: IEEE, 2018: 59−66

[39] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//Proc of the 38th Int Conf on Machine Learning. New York: ACM, 2021: 8748−8763

[40] Zhao Zhixian, Chen Haifeng, Li Xi, et al. Improving multimodal emotion recognition by leveraging acoustic adaptation and visual alignment[C]//Proc of the 2nd Int Workshop on Multimodal and Responsible Affective Computing. New York: ACM, 2024: 67−71

[41] Yang Aiyuan, Xiao Bin, Wang Bingning, et al. Baichuan 2: Open large-scale language models[J]. arXiv preprint, arXiv: 2309.10305, 2023

[42] Li Tianyi, Liu Daming. MPID: A modality-preserving and interaction-driven fusion network for multimodal sentiment analysis[C]//Proc of the 31st Int Conf on Computational Linguistics. Stroudsburg, PA: ACL, 2025: 4313−4322

[43] Xin Feng, Gong Wenjuan, Shi Tingbo, et al. Gated fusion network with progressive modality enhancement for sentiment analysis[J]. IEEE Access, 2024, 12: 165810−165821

[44] Fan Qi, Li Yong, Xin Yi, et al. Leveraging contrastive learning and

self-training for multimodal emotion recognition with limited labeled samples[C]//Proc of the 2nd Int Workshop on Multimodal and Responsible Affective Computing. New York: ACM, 2024: 72−77

[45] Gao Fei, Shi Pujin. Audio-Guided fusion techniques for multimodal emotion analysis[C]//Proc of the 2nd Int Workshop on Multimodal and Responsible Affective Computing. New York: ACM, 2024: 62−66

[46] Lei Ling, He Wangjun, Zheng Qiuyan, et al. Multimodal sentiment analysis based on expert mixing of subtask representations[J]. IEEE Access, 2025, 13: 44278−44287

[47] Liu Yihe, Yuan Ziqi, Mao Huisheng, et al. Make acoustic and visual cues matter: CH-SIMS v2. 0 dataset and av-mixup consistent module[C]//Proc of the 24th Int Conf on Multimodal Interaction. New York: ACM, 2022: 247−258



Chen Haifeng, born in 1988. PhD, lecturer. Member of CCF. His main research interests include affective computing, multimodal emotion recognition, emotion recognition in conversations, multimodal depression analysis, and facial action unit recognition.

陈海丰, 1988 年生。博士，讲师。CCF 会员。主要研究方向为情感计算、多模态情感识别、对话中的情感识别、多模态抑郁症分析、面部动作单元识别。



Li Xi, born in 2001. Master candidate. Student member of CCF. Her main research interest includes multimodal emotion recognition.

李　溪, 2001 年生。硕士研究生。CCF 学生会员。主要研究方向为多模态情感识别。



Li Yan, born in 1998. Master candidate. His main research interests include machine learning and multimodal understanding.

李　岩, 1998 年生。硕士研究生。主要研究方向为机器学习、多模态理解。



Li Jian, born in 1975. PhD, professor. His main research interests include computer vision and digital twin.

李　健, 1975 年生。博士，教授。主要研究方向为计算机视觉、数字孪生。



He Lang, born in 1983. PhD, associate professor. His main research interests include multimodal depression recognition and analysis, discrete emotion analysis, dimensional emotion recognition, and image processing.

何　浪, 1983 年生。博士，副教授。主要研究方向为多模态抑郁症识别和分析、离散情感分析、维度情感识别、图像处理。



Jiang Dongmei, born in 1973. PhD, professor. Her main research interests include audio-visual fusion based speech processing, audio-visual emotion analysis and recognition, and facial animation synthesis of speakers with expressions.

蒋冬梅, 1973 年生。博士，教授。主要研究方向为听视觉融合的语音处理、听视觉情感分析与识别、带表情的说话人面部动画合成。