基于知识蒸馏与动态调整机制的多模态情感分析模型

王 楠[1),2)] 王 淇[1)] 欧阳丹彤[3)]

[1)](吉林财经大学管理科学与信息工程学院 长春 130117)

[2)](吉林财经大学大数据与交叉科学研究院 长春 130117)

[3)](吉林大学计算机科学与技术学院 长春 130012)

摘 要 近年来,模态缺失已成为多模态情感分析中的重要挑战。然而,现有研究无法有效应对模态缺失场景,导致模型性能显著下降。为解决这一问题,本文提出了基于知识蒸馏与动态调整机制的多模态情感分析模型(Attention-based Uncertain Missing Modality Distillation Framework, AUMDF)。具体而言,设计了一种模态随机缺失策略,以增强模型对不确定模态场景的适应能力。此外,引入了动态权重调整模块和多模态掩码Transformer,用于平衡特征贡献并捕获模态间的细微交互。最后,设计了对比样本蒸馏和基于相似性的表示蒸馏机制,以加强教师模型与学生模型之间的对齐,实现高效的知识传递。在两个基准数据集上的实验结果表明,本文利用知识蒸馏和动态调整机制实现了对多模态数据之间模态交互关系的充分利用,并弥补了模态缺失场景下的研究缺陷。与现有的先进方法相比,在CMU-MOSI数据集上,AUMDF将平均绝对误差降低了0.8%,F1得分提高了0.3%;在CMU-MOSEI数据集上,AUMDF将平均绝对误差降低了0.2%,F1得分提高了0.3%;在IEMOCAP数据集上,AUMDF在“悲伤”与“愤怒”的情感分类中将F1得分分别提高了0.7%和0.2%。

关键词 知识蒸馏;多模态情感分析;注意力机制;多模态学习;特征融合

中图法分类号 TP18 **DOI号** 10.11897/SP.J.1016.2025.01923

Multimodal Sentiment Analysis Model Based on Knowledge Distillation and Dynamic Adjustment Mechanism

WANG Nan[1),2)] WANG Qi[1)] OUYANG Dan-Tong[3)]

[1)](Department of Management Science and Information Engineering, Jilin University of Finance and Economics, Changchun 130117)

[2)](Institute of Big Data and Interdisciplinary Sciences, Jilin University of Finance and Economics, Changchun 130117)

[3)](Department of Computer Science and Technology, Jilin University, Changchun 130012)

Abstract Multimodal Sentiment Analysis (MSA) has gained increasing attention due to its ability to leverage complementary information from textual, acoustic, and visual modalities, enabling more accurate and nuanced emotional inference compared to unimodal approaches. Despite its potential, the practical deployment of MSA models remains severely constrained by the problem of modal missingness, where one or more modalities may be absent due to factors such as sensor failures, data corruption, transmission loss, or privacy restrictions. In real-world applications, this issue is particularly prominent in scenarios like video-based sentiment analysis, where occluded facial expressions, poor lighting conditions, or degraded audio signals can significantly

收稿日期:2024-12-12;在线发布日期:2025-04-21。本课题得到国家社会科学基金项目(22BTQ048)资助。 **王 楠**,博士,教授,中国计算机学会(CCF)会员,主要研究领域为机器学习、自然语言处理、多模态学习。E-mail:119014@jlufe.edu.cn。 **王 淇**,硕士研究生,主要研究领域为多模态学习。**欧阳丹彤**(通信作者),博士,教授,主要研究领域为基于模型的诊断与自动推理。E-mail:ouyd@jlu.edu.cn。

disrupt emotional cues, leading to substantial performance degradation. To mitigate these challenges, existing approaches primarily adopt three types of strategies: modality dropout-based data augmentation, generative adversarial network (GAN)-based data imputation, and joint learning frameworks with shared latent representations. While data augmentation techniques enhance a model's adaptability to missing modalities, they fail to explicitly capture cross-modal dependencies and often introduce inconsistencies in representation learning. GAN-based approaches attempt to generate missing modalities but frequently suffer from distribution mismatches, excessive computational overhead, and instability in adversarial training. Meanwhile, joint learning frameworks, which either concatenate modality-specific features or enforce a shared latent space, struggle to model deep cross-modal interactions, resulting in limited robustness when facing highly sparse or irregularly missing modalities. Given these challenges, there remains a critical need for a more effective framework that not only dynamically adapts to missing modalities but also fully exploits cross-modal interactions for robust sentiment prediction. To address these challenges, this paper proposes an Attention-based Uncertain Missing Modality Distillation Framework (AUMDF) based on knowledge distillation and dynamic adjustment mechanism. AUMDF consists of three key components: (1) Dynamic Cross-Modal Weight Adjustment, a modality-aware attention mechanism that dynamically recalibrates feature contributions based on the presence and reliability of each modality, ensuring adaptive fusion even under missing conditions by prioritizing informative signals while mitigating noise from unreliable sources; (2) Multimodal Masked Transformer, a transformer-based encoder that explicitly models cross-modal dependencies through masked self-attention, enabling the model to infer missing information from available cues and enhance robustness to incomplete data; and (3) Dual-Distillation Learning, which combines contrastive sample distillation and similarity-based representation distillation to transfer knowledge from a teacher model (trained on complete data) to a student model (adapted to missing modalities), ensuring consistent and robust predictions while preserving the teacher model's discriminative power and adapting it to the student's domain for improved generalization under modality sparsity. Extensive experiments on benchmark datasets—CMU-MOSI, CMU-MOSEI, and IEMOCAP—demonstrate the effectiveness of AUMDF, achieving a 0.8% reduction in mean absolute error (MAE) and a 0.3% improvement in F1-score on CMU-MOSI, a 0.2% decrease in MAE and a 0.3% increase in F1-score on CMU-MOSEI, and 0.7% and 0.2% F1-score gains for "sadness" and "anger" on IEMOCAP, respectively. These results highlight AUMDF's capability to exploit cross-modal synergies and maintain predictive stability under modality sparsity. By explicitly addressing uncertain modal missingness through dynamic adaptation and knowledge distillation, AUMDF advances the state of the art in MSA, offering a robust and practical solution for real-world applications. Future work will explore extending AUMDF to additional modalities and integrating online learning paradigms for adaptive deployment in dynamic environments.

Keywords knowledge distillation; multimodal sentiment analysis; attention mechanism; multimodal learning; feature fusion

1 引 言

近年来,随着微博、Facebook和Twitter等在线社交平台的迅速普及,用户生成数据的规模以指数级增长。这些数据通常包括文本、图像、音频和视频等多种模态信息,用户利用这些模态资源表达个人观点、分享情绪体验或参与社会互动。这一趋势不

仅加速了信息传播的速度和范围，还使情感表达形式变得更加多样化和复杂化。因此，如何从多模态数据中有效提取和分析用户的情感信息，成为当前自然语言处理领域的重要研究方向之一。这一任务被称为多模态情感分析（Multimodal Sentiment Analysis，MSA），旨在通过集成多模态信息的互补特性，实现对用户情绪状态的自动识别和深度理解[1]。与传统单模态情感分析相比，多模态情感分析同时利用语言、音频、图像和视频等多种数据源，能够捕捉更加细腻的情感变化，提供更加全面、细致的情感解读。

在MSA的现实场景中，由于数据采集设备故障、传输延迟或用户隐私保护需求等原因，模态数据缺失问题普遍存在。然而，目前大多数研究仍然假设所有模态数据始终可用，导致现有模型的适应性和鲁棒性不足。因此，如何在动态模态缺失条件下有效融合多模态信息，成为当前多模态情感分析研究中的重要挑战之一。

为解决上述问题，本文提出一种基于知识蒸馏与动态调整机制的多模态情感分析模型（Attention-based Uncertain Missing Modality Distillation Framework，AUMDF）。该框架以模态动态缺失环境下的情感分析任务为目标，通过模拟模态缺失场景、设计自适应优化特征表示，在“教师-学生”模型架构中实现知识传递，从而提升模型在复杂环境下的情感识别能力。

AUMDF的核心目标和具体贡献如下：

（1）建立一种模态缺失场景的知识蒸馏学习机制。在训练阶段，AUMDF首先采用模态随机丢失策略在学生模型端模拟真实场景中的动态模态缺失情况，然后通过设计对比样本蒸馏（CSD）与基于相似性的表示蒸馏（SRD）两种策略，在教师模型和学生模型之间建立高效的知识对齐机制，实现知识的有效传递，使其能够在不完整数据环境下依然保持较高的情感分析准确率，克服了现有知识蒸馏方法仅关注输出空间的软目标对齐，而忽视特征空间中深层知识传递的局限性。

（2）特征增强处理与跨模态信息挖掘。为了挖掘不同模态特征之间的协同信息，AUMDF分别设计了特征增强模块（REM）、特征预处理模块（RPM）和动态权重调整模块（DWAM）对特征进行优化与融合。首先，REM通过异质信息统一处理和噪声过滤两个子模块，有效保留输入特征的全局信息，并增强去噪后的局部信息表达能力；接下来，RPM通过一维卷积操作同时融入位置信息，独立处理每种模态以有效捕捉其独特特征；最后，DWAM引入门控机制融合跨模态的显著特征信息，确保最具信息量的模态得到更强的关注，充分挖掘模态间的相关性。

（3）跨模态特征的细粒度深层交互。针对传统Transformer难以充分捕捉多模态间全局交互关系的缺点，本文提出多模态掩码Transformer，结合模态间相似度计算挖掘跨模态之间的语义关联，识别关键语义线索，捕捉模态内部与跨模态之间的情感信息细粒度交互关系，为情感预测提供更加精准的特征支持。

2　相关工作

2.1　多模态情感分析

多模态情感分析作为情感计算领域的重要研究方向，近年来得到了广泛关注[2]。与传统的单一模态情感分析方法相比，多模态情感分析通过融合来自视频、音频、文本等多种模态的信息，能够显著提高情感分析的准确性和鲁棒性。

在MSA研究的初期阶段，传统的机器学习方法得到广泛应用，并且在一定程度上取得了成功。典型的研究方法如Rozgić等人[3]提出的基于支持向量机（SVM）分类器的自动生成树模型，通过对多模态数据进行特征提取和分类，成功地实现了情感识别；Cummins等人[4]通过引入词袋模型和跨领域数据，进一步提高了情感检测系统的性能；Wang等人[5]则提出了一种面向微博图像的跨媒体词袋方法，联合使用图像和文本特征，并采用逻辑回归、SVM和朴素贝叶斯等经典分类器进行训练，显著提升了情感分类效果。这些传统方法的优势在于其较为简单的实现方式和相对较低的计算成本，但也存在无法深入挖掘模态间复杂关系的局限性。

多模态情感分析的关键在于如何有效地将不同模态的信息进行融合，以便提取出更加精细和全面的情感特征。因此，如何设计出高效的多模态特征融合策略，成为了当前研究的核心问题[6-8]。随着深度学习技术的飞速发展，基于深度学习的多模态情感分析方法逐渐成为主流，这些方法主要基于卷积神经网络（Convolutional Neural Network，CNN）、长短时记忆网络（Long Short-Term Memory，LSTM）和门控循环单元（Gated Recurrent Unit，GRU）等深度学习模型[9-12]，这些模型已经在多种任

务上取得了很好的结果[13-16]，并在情感分析任务中表现出了显著的性能提升[17-18]。深度学习模型能够自动从多模态数据中学习复杂的特征，并在多个层次上进行表示，进而提取更加丰富的情感信息。根据融合多模态信息的时机和方式，深度学习方法可分为早期融合和后期融合两类。早期融合方法通过在输入阶段对多模态信息进行整合，以生成统一的情感表示[19-21]。这种方法能够较好地捕捉模态间的相关性，并生成一组共享的特征表示。然而，早期融合的方法也存在一定的不足，特别是当某一模态的质量较差时，整个模型的性能容易受到影响。与此不同，后期融合方法则是先对各个模态独立进行情感分析，然后再将各个模态的预测结果进行融合，从而生成最终的情感判断[22-24]。后期融合方法能够保持各模态的独立性，减少个别模态信息缺失对整体情感分析结果的影响，但其局限在于模态间的潜在交互信息未能得到有效利用，导致情感分析的精度无法得到充分提升。

尽管早期融合和后期融合方法在多模态情感分析中取得了一定的进展，但它们在处理模态间差异性和模态间交互信息的利用上仍存在不足。一些研究尝试根据不同模态对情感分析任务的贡献动态地调整模态的权重。例如，罗渊贻等人[25]提出一种多模态学习方法，通过提高各模态的共性特征表达的方式增加不同模态间的动态交互以提高语义一致性；Arevalo等人[26]提出了一种基于门控网络的权重分配方法，通过动态加权不同模态的信息，以提升融合效果；Yang等人[27]则提出了一种自适应路径选择机制，为每个模态分配不同的权重，从而实现更加精细的模态融合。这些方法通过调节权重来突出不同模态的贡献，增强了情感分析的表现。除上述方法外，一些学者还提出基于图融合方法进行情感的识别，例如，宗林林等人[28]提出基于超图的多模态情绪识别模型，通过引入超图建立多模态的多元关系，以此替代现有图结构采用的多个二元关系，实现了更加充分、高效的多模态特征融合。然而，这些方法仍然存在对模态间交互信息挖掘不深的问题，导致在情感分析任务中融合效果不够充分。

近年来，注意力机制的引入为多模态情感分析提供了新的思路，并取得了显著的研究进展。与基于权重调节的方法不同，注意力机制能够更加灵活地根据模态对情感任务的贡献动态调整特征权重，同时有效建立模态间的上下文关系，解决了传统方法中对模态间信息挖掘不足的问题。如Zhang等人[29]提出的全局注意力模块，能够通过学习全局上下文和局部细节的注意力权重，从而增强对多模态数据的感知能力；Lian等人[30]利用双向GRU和多重注意力框架，探索了多模态对话中的情感交互，通过深度建模模态间的相互关系，提升了情感分析的精度。此外，Mai等人[31]设计的基于模态内和模态间的对比学习方法，通过对模态间关系的建模，进一步提升了跨模态交互特征的捕捉能力。Ashima等人[32]提出的双重注意力网络（DMLANet）则通过生成视觉和语言间的双向注意力图，进一步提高了情感分类任务的细粒度能力。这些方法在一定程度上提升了情感分析的表现，尤其在模态间交互信息的挖掘方面，取得了显著进展。

在现实应用中，由于传感器或输入数据的不稳定性，某些模态可能会出现缺失，从而影响整体的情感分析效果。一些学者针对MSA中的这种模态缺失问题展开了研究[33-43]。现有模态缺失补充方法主要分为基于特征重建的方法以及基于图结构与对比学习的方法。在基于特征重建的方法中，MCTN[38]使用从原模态目标的循环转移来训练联合表示，确保模型对缺失模态场景建模的鲁棒性；Sun等人[40]提出了一种双层特征重建机制，通过低级特征重建隐式引导模型从不完整数据中学习语义信息；Shi等人[41]提出的TgRN模型进一步引入了文本引导（text-guided）策略，通过文本模态对非文本模态进行特征重建，并在非对齐序列上优化模态缺失下的信息融合过程。在基于图结构与对比学习的方法中，文献[42]提出一种专门用于对话场景的模态缺失学习框架，将不完整对话数据看作部分缺失的图结构，并通过端到端的分类和重建联合优化机制，提升模型在模态缺失条件下的对话情感识别能力。然而，大多数现有的多模态情感分析模型仍然假设所有模态在进行情感分析时都是可用的。这种假设在面对模态缺失的情境时，往往导致模型性能的显著下降。因此，如何有效应对模态缺失问题，依旧是当前多模态情感分析领域的重要研究挑战。

2.2 知识蒸馏

知识蒸馏（Knowledge Distillation，KD）作为一种模型压缩和迁移学习的方法，旨在通过减少教师模型（Teacher Network）与学生模型（Student Network）之间的输出分布差异，将教师模型的知识有效地传递给学生模型[44]。知识蒸馏的核心思想是通过教师模型的“软标签”（Soft Target）引导学生模型的学习，从而提升学生模型的泛化能力，该方法已

经被应用到各种多模态数据的研究中。如在多模态数据处理中，Wei等人[45]提出了一种新型的层次化跨模态交互与融合网络，该网络通过自蒸馏机制在多个层次上进行知识传递，从而弥合不同模态间的语义鸿沟；Zhang等人[46]设计了一种样本加权蒸馏与原型正则化网络，通过优化样本权重来解决模态缺失和不平衡问题。尽管这些方法取得了一定成果，但它们仍然存在一定的局限性，如未能充分捕捉模态间的细粒度交互信息，或者在知识传递过程中出现信息丢失等。

在多模态情感分析任务中，知识蒸馏也已被成功应用于解决模态缺失问题[47-49]。教师模型对完整模态数据进行训练，学生模型则在缺失模态的情况下，通过学习教师模型的预测结果，克服模态缺失带来的负面影响。例如，Sun等人[50]提出了一种基于补全的模型，通过最相似的模态数据填补缺失模态，实现知识的有效转移，增强模型在模态缺失情况下的鲁棒性；Wei等人[51]则提出了边缘感知蒸馏和模态感知正则化方法，旨在通过提升强模态和弱模态的表示能力，帮助模型应对模态缺失问题；Deng等人[52]引入了一种基于图蒸馏的框架，利用图结构填充缺失数据，并实现多模态数据的聚合，从而有效地处理模态缺失带来的挑战。

综上所述，尽管知识蒸馏技术在多模态情感分析任务中取得了一定的进展，但在面对模态缺失和模态间细粒度交互问题时，仍然存在一定的挑战。为了解决这些问题，本文提出融合动态调整模块与多模态掩码Transformer的多模态情感分析模型（AUMDF），该模型不仅能够有效应对模态缺失问题，还能通过加强模态间的相互学习，进一步提升情感分析的性能。

3　融合动态调整模块与多模态掩码Transformer的多模态情感分析模型

3.1　问题定义

多模态情感分析通常被定义为回归任务，目标是预测语句级情感分数。给定包含完整模态的样本集$S=[\boldsymbol{X}_L,\boldsymbol{X}_A,\boldsymbol{X}_V]$，其中$\boldsymbol{X}_L\in\mathbb{R}^{T_L*d_L}$、$\boldsymbol{X}_A\in\mathbb{R}^{T_A*d_A}$和$\boldsymbol{X}_V\in\mathbb{R}^{T_V*d_V}$分别表示文本、视觉及音频的单峰特征，本文后续使用$m$代表模态种类，用$t$代表教师模型，用$s$表示学生模型，如$X_L^t$为教师模型输入的文本模态数据。$T_m$为模态$m(m\in\{L,A,V\})$的序列长度，$d_m$为嵌入维度。为进一步拓展传统的MSA任务，本文提出将其扩展至模态缺失场景。在该场景下，部分模态的特征缺失，记作$\boldsymbol{X}_m^s$，目标是使得模型在模态缺失情况下也能够预测情感分数。

3.2　整体结构

为解决多模态情感分析中的模态缺失问题，本文设计了一个融合动态调整模块与多模态掩码Transformer的多模态情感分析模型（AUMDF）。该框架包括教师模型和学生模型，二者具有相同的结构，但参数不同。图1展示了AUMDF的整体工作流程。

在训练阶段，首先使用完整模态样本集S训练教师模型，并冻结其参数以保持模型稳定性。其次，通过模态随机缺失策略生成缺失模态样本集$\hat{S}=[\boldsymbol{X}_L^s,\boldsymbol{X}_A^s,\boldsymbol{X}_V^s]$，其中缺失模态的特征部分将用零向量进行填充。最后，将完整样本集S输入教师模型，同时将缺失模态样本集$\hat{S}$输入学生模型。

为了使模态特征能够适配后续的融合过程，本文通过特征增强模块与特征预处理模块对输入样本S和$\hat{S}$进行转换，生成对应的模态特征$\hat{\boldsymbol{Z}}_m^t$和$\hat{\boldsymbol{Z}}_m^s$，为后续的模态特征融合做准备。接下来，进一步设计了动态权重调整模块和多模态掩码Transformer来融合不同模态的特征，生成教师模型和学生模型的联合表示$\boldsymbol{H}^t$和$\boldsymbol{H}^s$。

在模态特征融合之后，本文在训练过程中引入了两种基于知识蒸馏的特征训练策略，包括对比样本蒸馏（CSD）和基于相似性的表示蒸馏（SRD）。CSD机制通过对比正负样本对的相似度，强化学生模型在缺失模态情境下的学习能力；而SRD机制则通过将教师模型和学生模型的表示进行对齐，提升其一致性与互补性，从而增强模型的预测能力。

在推理阶段，只有学生模型输出的融合特征$\boldsymbol{H}^s$被输入到任务特定的分类器中，生成最终的情感预测结果$\hat{y}^s$，以完成情感分析任务。

3.3　模态随机丢失策略(RPRM)

为了模拟真实场景中的动态模态缺失情况，本文提出了模态内随机缺失策略，通过随机丢弃模态的帧级特征来生成缺失模态数据集，作为学生模型端的输入。对于模态内损失，本文定义一个丢弃率p，取值范围设定为[0,0.7]，步长为0.1。丢失率p针对每个模态应用，即每个模态的数据可能丢失一定比例的特征，但仍然保留一定数量的数据。例如，当丢弃率$p=0.5$时，每个模态将随机丢弃50%的数据。不同于以往的研究[53]，本文并未在$p=1.0$

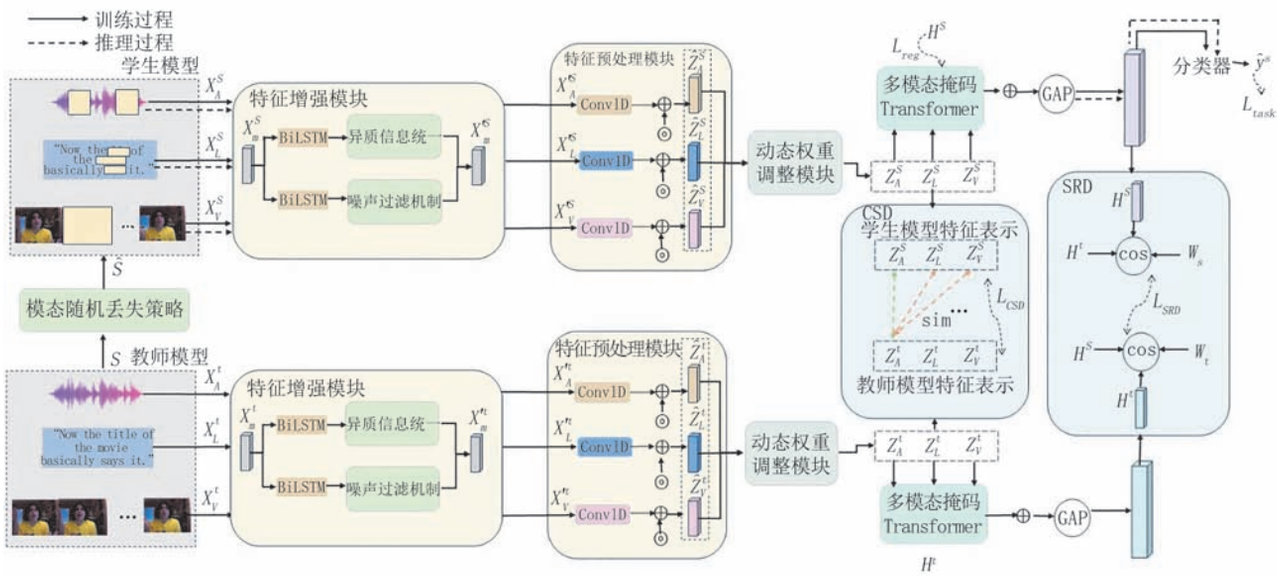

图1 AUMDF 结构框架

时进行实验，因为完全丢失所有模态的数据将失去实验的意义。缺失的部分将被填充为零向量，从而生成不完整的模态特征 $\boldsymbol{X}_m^s(m\in\{L,A,V\})$。

这一策略能够有效增强模型在处理缺失模态时的鲁棒性，使得即使某些模态数据缺失，模型仍能从其他模态中提取有效信息，从而提升情感分析的性能。

3.4 特征增强模块(REM)

多模态情感分析的关键挑战在于如何充分挖掘不同模态特征之间的协同信息，同时抑制噪声干扰和冗余特征。特征表示增强模块的主要作用在于规范化输入特征、提升特征表示的有效性以及加强模态间的语义关联性，包含两个子模块，分别为异质信息统一子模块和噪声过滤子模块。

3.4.1 异质信息统一子模块

异质信息统一子模块主要针对包含文本、视觉与音频的多模态输入特征提取统一表示。不同模态数据的表达方式各具特点，例如文本通常通过时序数据呈现，而图像则通过像素的空间分布表达。因此，为了保证这些特征可在后续步骤中进行有效融合，需先对不同模态中的异质信息进行统一的特征建模。

首先，接收到各模态单峰特征之后，引入双向长短时记忆网络(Bidirectional Long Short-Term Memory，BiLSTM)来捕获全局的时序依赖关系，同时考虑序列的前向和后向信息，从而更全面地编码上下文关系，具体过程可表示为：

$$\bar{\boldsymbol{X}}_m=\left[x_1^m,x_2^m,\cdots,x_{T_m}^m\right]\in\mathbb{R}^{T_m*d_m} \tag{1}$$

$$x_h^m=BiLSTM\left(\bar{\boldsymbol{X}}_m;\theta_h^m\right),\forall h\in\{1,2,\cdots,T_m\} \tag{2}$$

其中，$\bar{\boldsymbol{X}}_m$ 表示模态 m 的输入序列，x_h^m 表示在时间步长 h 时刻的隐藏状态，θ_h^m 为模型参数。通过协同提取前向和后向的时序特征，模型能够捕捉到输入特征在时序维度上的完整语义信息。

获取上述时序特征后，为消除不同模态之间的尺度差异，进一步进行标准化处理，这一过程可表示为

$$\hat{x}_h^m=\frac{x_h^m-\mu^m}{\sqrt{\sigma^m+\varphi^m}} \tag{3}$$

其中，μ^m、σ^m 分别表示模态 m 特征的均值与平均差，φ^m 表示模态 m 特征的平滑因子，用以防止除零错误。标准化后的特征不仅在数值范围上保持一致，也提升了整个训练过程中的稳定性，以促进不同模态特征间的融合。μ^m、σ^m 的具体计算过程可表示为

$$\mu^m=\frac{1}{T_m}\sum_{h=1}^{T_m}x_h^m \tag{4}$$

$$\sigma^m=\frac{1}{T_m}\sum_{h=1}^{T_m}\left(x_h^m-\mu^m\right)^2 \tag{5}$$

$$\hat{\boldsymbol{X}}_h^m=\left[\hat{x}_1^m,\hat{x}_2^m,\cdots,\hat{x}_{T_m}^m\right] \tag{6}$$

3.4.2 噪声过滤子模块

注意力机制可以通过特征加权方式达到一定的噪声过滤效果，但最终只是降低了噪声特征的权重，噪声特征本身仍然被保留，为了得到更加纯净的多模态输入，本文提出噪声过滤子模块。该模块通过

设计非线性激活和门控机制，完全剔除了低于阈值的特征，并引入噪声补偿特征，在抑制噪声的同时为缺失部分提供合理补偿，从而显著提升了特征输入的纯净性，更有效地抑制了噪声干扰。

首先，与异质特征处理模块同样引入双向长短时记忆网络（BiLSTM）进行时序特征提取，接着应用ReLU激活函数进行非线性变换，获取这一模块的激活向量$\hat{x}_f^m$，具体计算过程如公式（7）和公式（8）所示：

$$x_f^m = BiLSTM\left(\bar{\boldsymbol{X}}_m;\ \theta_f^m\right), \forall f \in \{1, 2, \cdots, T_m\} \tag{7}$$

$$\hat{\boldsymbol{x}}_f^m = ReLU\left(W_f^m x_f^m + b_f^m\right) \tag{8}$$

其中，W_f^m与b_f^m分别表示模态m特征的权重矩阵与偏置项。完成非线性变换后，本文进一步引入根据特定阈值过滤特征的动态门控机制，可表示为

$$Z_f^m = \begin{cases} 1, & \text{if } x_f^m \geqslant \rho^m \\ 0, & \text{if } x_f^m < \rho^m \end{cases} \tag{9}$$

其中，ρ^m表示模态m特征的噪声过滤阈值。通过门控机制，保留有效特征后显著提高特征输入的质量，减少不相关信息对下游情感分析任务的干扰。其次，将门控机制Z_f^m应用于激活向量$\hat{\boldsymbol{x}}_f^m$，生成去噪特征$\tilde{\boldsymbol{x}}_f^m$，具体计算过程可表示为

$$\tilde{\boldsymbol{x}}_f^m = Z_f^m * \hat{\boldsymbol{x}}_f^m + \left(1 - Z_f^m\right) * \boldsymbol{\delta}^m \tag{10}$$

其中，$\boldsymbol{\delta}^m$表示模态m特征的噪声补偿特征，其通过拼接模态m特征的全局均值与标准差，再输入至MLP层处理得到。$\boldsymbol{\delta}^m$充分利用了模态特征的全局统计信息，为缺失或低质量特征提供合理补偿，减少了噪声信息对最终预测结果的影响。

最后，为了在异质信息统一特征$\hat{\boldsymbol{X}}_h^m$和去噪特征$\tilde{\boldsymbol{x}}_f^m$之间实现动态平衡，本文设计动态权重融合机制生成最终的融合特征$\boldsymbol{X'}_m^t$，具体计算方法表示如下：

$$\tilde{\boldsymbol{X}}_f^m = \left[\tilde{\boldsymbol{x}}_1^m, \tilde{\boldsymbol{x}}_2^m, \cdots, \tilde{\boldsymbol{x}}_{T_m}^m\right] \tag{11}$$

$$\boldsymbol{X'}_m^t = \alpha^m * \hat{\boldsymbol{X}}_h^m \otimes \left(1 - \alpha^m\right) * \tilde{\boldsymbol{X}}_f^m \tag{12}$$

其中，α^m表示模态m特征的动态平衡参数，用以调节异质信息统一特征和去噪特征之间的权重比例。

通过上述过程，AUMDF能够有效保留异质信息统一化之后的全局特征信息，同时增强去噪后的局部信息表达能力。

3.5　特征预处理模块（RPM）

为了独立处理每种模态并有效捕捉其独特特征，首先将每种模态的输入表示为一个标记序列。假设给定模态m（$m \in \{L, A, V\}$）的特征序列$\boldsymbol{X}_m \in \mathbb{R}^{T * d_m}$，其中$T$表示序列的长度，$d_m$表示模态$m$的特征维度。现有研究（如文献[31]）通常直接采用原始特征形式输入模型，忽略了时间序列的依赖关系，无法有效捕捉各模态独特特征。为了有效捕捉时间序列中的局部依赖关系，本文采用一维卷积操作对每种模态的输入特征进行处理，从而生成模态特征的嵌入表示$\boldsymbol{X}_m^{t\triangle T}$，如公式（13）所示：

$$\hat{\boldsymbol{X}}_m^{t\triangle T} = Conv1D\left(\boldsymbol{X}_m^t, k_m^t;\ \theta_{\triangle T}\right) \tag{13}$$

其中，k_m^t表示卷积核的大小，$\theta_{\triangle T}$表示一维卷积网络在时间步$\triangle T$时刻的参数。通过该卷积过程，模型能够捕捉语言模态中的相邻单词之间的关系，或者音频模态中的连续依赖，从而保留每种模态内部的时序特征。

进一步地，为了增强模型对序列元素位置信息的感知，本文在特征向量中加入了位置编码，具体定义如公式（14）和公式（15）所示：

$$PE\left(pos, 2k\right) = \sin\left(\frac{pos}{10\,000^{\frac{2k}{d}}}\right) \tag{14}$$

$$PE\left(pos, 2k+1\right) = \cos\left(\frac{pos}{10\,000^{\frac{2k}{d}}}\right) \tag{15}$$

其中，PE表示序列中每个元素的位置信息，d为特征的维度。通过在特征向量中添加位置编码，模型能够在后续的处理过程中有效保留时间序列的顺序信息，确保序列输入的时间依赖性得到正确捕捉。最终，模态特征的嵌入表示通过以下方式计算，即

$$\tilde{\boldsymbol{Z}}_m^{t\triangle T} = \hat{\boldsymbol{X}}_m^{t\triangle T} + PE(T, d) \in \mathbb{R}^{T * d} \tag{16}$$

其中，$\tilde{\boldsymbol{Z}}_m^{t\triangle T}$为模态$m$在时间步$\triangle T$时刻的最终嵌入表示。通过卷积操作捕捉局部依赖，并结合位置编码融入位置信息，确保了每种模态的特征在后续多模态融合阶段的准确性和一致性。

3.6　动态权重调整模块（DWAM）

如图2所示，不同于仅依赖文本模态作为指导特征的方法[54-55]，动态权重调整模块旨在通过融合其它模态的特征并引入动态权重机制来优化原始表示空间，从而实现跨模态的深度融合。相较于基于单模态内部上下文分配权重关系的自注意力机制，动态权重调整模块引入门控机制，通过显式建模模态间的协同关系，能够确保最具信息量的模态得到更强的关注，使得模型能够充分挖掘模态间的相关性，进而提升情感预测的准确性。

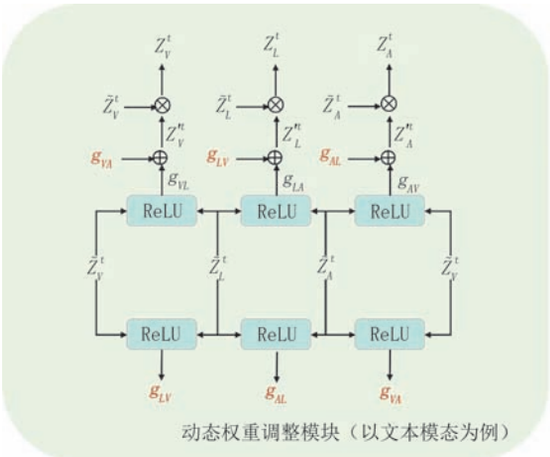

图2 动态权重调整模块架构

对于教师模型中的动态权重调整模块，输入由特征预处理模块输出的嵌入表示$\tilde{Z}_m^t$。首先，为了获得各模态的门控向量，本文将预处理后的文本表示$\tilde{Z}_L^t$、音频表示$\tilde{Z}_A^t$和视频表示$\tilde{Z}_V^t$进行拼接，并通过ReLU激活函数生成相应的门控向量。以生成文本门控向量的过程为例，具体计算如公式(17)与公式(18)所示：

$$g_{LA}=ReLU\left(\boldsymbol{W}_{LA}\left[\tilde{Z}_L^t;\tilde{Z}_A^t\right]+\boldsymbol{b}_{LA}\right) \tag{17}$$

$$g_{LV}=ReLU\left(\boldsymbol{W}_{LV}\left[\tilde{Z}_L^t;\tilde{Z}_V^t\right]+\boldsymbol{b}_{LV}\right) \tag{18}$$

其中，$\boldsymbol{W}_{LA}$和$\boldsymbol{W}_{LV}$是权重矩阵，b_{LA}和b_{LV}为偏置项。通过此门控机制，模型能够为每个模态生成不同的权重，从而有选择性地加强某些模态的信息贡献。

随后，本文将生成的文本门控向量g_{LA}和g_{LV}进行融合，得到加权后的文本特征表示Z_L^t，如公式(19)与公式(20)所示：

$$Z'^t_L=\left[g_{LA};g_{LV}\right] \tag{19}$$

$$Z_L^t=Z'^t_L\cdot\left(\boldsymbol{W}_{LW}\tilde{Z}_L^t\right)+\boldsymbol{b}_{LW} \tag{20}$$

同样的步骤也被应用于音频和视觉模态，以得到加权后的音频表示Z_A^t和加权后的视觉表示Z_V^t。通过融合跨模态的显著特征信息，动态权重调整模块可以为后续的情感预测提供更加全面和准确的模态表示。

3.7 多模态掩码Transformer(MMT)

经过DWAM的处理后，模型能够自适应地对不同模态特征给予不同程度的关注，强化了信息量丰富的模态特征在情感分析任务中的作用。然而，DWAM主要聚焦于单模态特征的独立增强，未能深入挖掘模态间潜在的语义关联与特征互补性。为解决这一问题，本文提出多模态掩码Transformer(MMT)，进一步对动态调整后的模态特征进行深层次交互与一致性建模。

首先，与多数模型仅计算两两模态间注意力的方式相比，MMT通过构建以语言、音频和视觉模态为查询的三模态注意力机制，强化了特征交互与信息流动，能够捕捉更复杂的语义关联；其次，MMT通过引入时间窗口与时间权重，解决了如MulT[56]等现有模型存在的未显式处理多模态时序性对齐的问题，确保了跨模态交互在时序维度上的对齐；最后，通过引入掩码机制，MMT能够有效过滤冗余信息，提升模型在模态缺失场景下的鲁棒性。

如图3所示，多模态掩码Transformer通过将每个模态独立作为查询向量，整合来自其他模态的互补信息。该模块增强了跨模态特征的提取能力，并确保在模态缺失的情况下，模型依然能够进行稳健的情感表示学习。

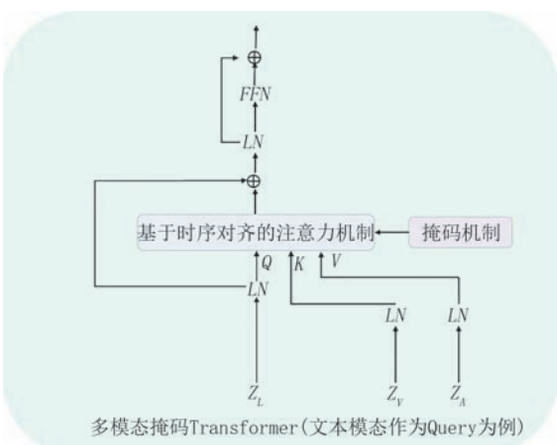

图3 多模态掩码Transformer架构

首先，为了处理不同模态中的时间依赖性，本文引入了时间对齐机制，具体计算过程可表示为

$$\Delta T_{ij}=\begin{cases}s(i,j), & \text{if}(i-j)\leqslant K\\ 0, & \text{otherwise}\end{cases} \tag{21}$$

其中，$s(i,j)$表示标记i和标记j之间的时间权重，K为预定义的时间窗口。通过时间对齐机制，确保来自不同模态的输入能够在适当的时间尺度内对齐，从而为后续的跨模态交互提供一致的时间信息。将时间信息整合到单峰表示中，具体计算方法如下所示：

$$Z_m^{t\Delta T}=Z_m^t+\Delta T_m \tag{22}$$

其次，采用掩码机制$Mask_m$，消除填充标记对注意力计算的影响，确保填充项不干扰模型的学习过

程。通过计算模态间的相似性，多模态掩码Transformer能够有效聚焦于与情感分析最相关的特征，从而提高情感预测的准确性，如公式(23)所示：

$$Mask_m=\begin{cases}0, & \text{if token on the padding}\\ 1, & \text{otherwise}\end{cases} \tag{23}$$

公式(23)处理输入序列中的padding token，确保它们不会对计算产生不必要的影响。通过这种方式，模型可以专注于非填充标记，避免填充标记对模态间相似性计算的干扰。

与传统注意力机制不同，本文基于模态特征的加权相似性，计算来自不同模态token之间的相似性进行对齐，即给定两个模态的特征向量后，通过点积计算来确定它们之间的相似性，能够有效突出对情感相关模态特征的聚焦，减少冗余信息的干扰，计算过程如公式(24)所示：

$$sim(\boldsymbol{\alpha},\boldsymbol{\beta})=\alpha_1\beta_1+\cdots+\alpha_d\beta_d \tag{24}$$

其中，d表示向量的维度。为了进一步扩展模态之间的相似性计算以得出模态间的注意力，本文引入了一个额外的d维模态向量$\boldsymbol{\gamma}$，并通过以下公式计算三模态之间的加权相似性：

$$sim(\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma})=\alpha_1\beta_1\gamma_1+\cdots+\alpha_d\beta_d\gamma_d \tag{25}$$

通过这一点积操作计算三种模态嵌入$\boldsymbol{\alpha}$、$\boldsymbol{\beta}$与$\boldsymbol{\gamma}$之间的加权相似度，模型能够捕捉不同模态之间的对齐关系，重点关注与情感相关的有效模态特征。

类似地，从向量转换为矩阵时，为加强相似性计算的表达能力，模型引入了Query、Key和Value三种向量，并引入权重矩阵进行动态调整，以便更准确提取各模态的情感特征，具体过程如公式(26)到公式(29)所示：

$$\boldsymbol{Q}_m=\boldsymbol{W}_Q^m\boldsymbol{Z}_m^{t\triangle T} \tag{26}$$

$$\boldsymbol{K}_m=\boldsymbol{W}_K^m\boldsymbol{Z}_m^{t\triangle T} \tag{27}$$

$$\boldsymbol{V}_m=\boldsymbol{W}_V^m\boldsymbol{Z}_m^{t\triangle T} \tag{28}$$

$$MMT(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=concat(head_1,\cdots,head_h)\boldsymbol{W}_O \tag{29}$$

其中，$\boldsymbol{W}_Q^m$、$\boldsymbol{W}_K^m$和$\boldsymbol{W}_V^m$分别表示查询、键和值的输入矩阵，$\boldsymbol{W}_O$是模型的可学习矩阵。通过这种多头注意力机制，模型能够从多个维度综合各模态的信息，并通过各个头的不同专注点来识别复杂的跨模态关系，从而提高情感预测的准确性。每个注意力头的具体计算过程可表示为

$$head_m=softmax\left(\frac{\boldsymbol{Q}_m{\boldsymbol{K}_m}^T}{\sqrt{d_k}}*Mask_m\right)\boldsymbol{V}_m \tag{30}$$

3.8　模型前向传播过程

与传统Transformer中一对一方式交互的跨模态注意力机制相比，本文提出的多模态掩码Transformer以三种模态的交互关系为核心，结合两两模态间的相似度权重与第三模态特征捕捉更复杂的跨模态语义关系，确保特征更加紧密对齐，以提升最终融合语义表示的质量。AUMDF的整体前向传播过程可按如下过程进行，首先，对于文本模态特征，其表示过程通过多模态掩码Transformer实现，如公式(31)和公式(32)所示：

$$\widehat{\boldsymbol{H}_L^t}=MMT\left(\boldsymbol{LN}(\boldsymbol{Z}_L^t),\boldsymbol{LN}(\boldsymbol{Z}_V^t),\boldsymbol{LN}(\boldsymbol{Z}_A^t)\right)+\boldsymbol{LN}(\boldsymbol{Z}_L^t) \tag{31}$$

$$\boldsymbol{H}_L^t=FFN\left(\boldsymbol{LN}\left(\widehat{\boldsymbol{H}_L^t}\right)\right)+\boldsymbol{LN}\left(\widehat{\boldsymbol{H}_L^t}\right) \tag{32}$$

其中，MMT表示多模态Transformer，FFN表示前馈网络。这种双重层归一化确保了特征在每个网络层中的稳定性，避免了信息的过度缩放或衰减。类似地，将多模态掩码Transformer应用于视觉特征与音频特征，可获得增强后的相应特征$\boldsymbol{H}_A^t$与$\boldsymbol{H}_V^t$。

接着，模型通过全局平均池化(GAP)来降低融合特征的维度。对于教师模型，这一操作通过平均化各模态特征值来保留模态间的关键信息，从而获得融合后的特征表示$\boldsymbol{H}^t$，具体计算过程如公式(33)所示：

$$\boldsymbol{H}^t=GAP\left(concat\left(\boldsymbol{H}_L^t,\boldsymbol{H}_V^t,\boldsymbol{H}_A^t\right)\right) \tag{33}$$

多模态数据通常容易遇到大量的类别内多样性和明显的类别间相似性，以往一些通过知识蒸馏解决模态缺失问题的方法[57-58]，多采用特征级融合方法，强调加强教师和学生模型之间的特征一致性。这种方法忽略了类别内复杂的相互作用，没有考虑潜在的特征变异性，易导致特征分布模糊和重叠。为了减轻这些限制，本文引入基于决策级融合策略的融合方式，在决策层对各模态特征进行整合，如公式(33)所示的教师模型端的融合表示$\boldsymbol{H}^t$，其汇集了各模态的情感信息，为最终的情感预测提供了全面的特征支持。对于学生模型，可通过相似操作获取融合表示$\boldsymbol{H}^s$。

在推理阶段，只输出学生模型的预测分数作为模型的最终输出。因此，将学生模型的融合表示输入至全连接层，结合可学习权重与Softmax函数输出预测的情感得分，如公式(34)所示：

$$\hat{y}^s=softmax\left(\boldsymbol{W}_{s\_pre}\boldsymbol{H}^s+\boldsymbol{b}_s\right) \tag{34}$$

3.9　对比样本蒸馏(CSD)

现有知识蒸馏方法[50,52]通常局限于单一样本层

面的知识迁移，未能有效建模跨样本的全局语义关联，导致学生模型在模态缺失场景下难以捕捉情感类别的共性特征与类间差异。针对这一缺陷，本文设计了对比样本蒸馏(CSD)机制，通过构建跨模态的正负样本对关系实现知识传递，促使模型学习到同一类别之间样本的深层情感关联，区分不同类别之间样本的情感特性。

本文将正样本集定义为与目标样本情感类别相同的样本集合，例如，如果目标样本属于“积极情感”类别，则所有标记为“积极情感”的其他样本均构成正样本集。通过将正样本的特征拉近，模型能够学习到同一情感类别样本的潜在相似性，从而提升表征的一致性。类似地，将负样本集定义为与目标样本情感类别不同的样本集合，通过将负样本的特征远离目标样本，使得模型能够区分不同情感类别样本的特性，明确类别边界并有效区分。具体而言，教师模型的文本模态对比损失函数定义为

$$L_L^t=-\frac{1}{|P_L|}\sum_{p\in P_L}\log\frac{exp\left(\alpha_{LV}*sim\left(\boldsymbol{X}_L^p,\boldsymbol{X}_V^s\right)+\alpha_{LA}*sim\left(\boldsymbol{X}_L^p,\boldsymbol{X}_A^s\right)\right)}{\sum_{s\in S_L}exp\left(\alpha_{LV}*sim\left(\boldsymbol{X}_L^p,\boldsymbol{X}_V^s\right)+\alpha_{LA}*sim\left(\boldsymbol{X}_L^p,\boldsymbol{X}_A^s\right)\right)} \tag{35}$$

其中，P_L表示文本模态的正样本集，S_L表示文本模态正负样本的全体集合，α_{LV}和α_{LA}表示可学习的用以调整文本-视觉样本对和文本-音频样本对相似性的缩放因子，其根据数据特征和模型学习过程自适应调整，在训练过程中通过反向传播自动优化，动态适应不同模态间的相似性分布。sim是计算模态间相似性的函数，其计算过程如公式(24)所示。

类似地，视觉模态与音频模态的对比损失也可以通过相同的方式定义，如公式(36)与公式(37)所示：

$$L_V^t=-\frac{1}{|P_V|}\sum_{p\in P_V}\log\frac{exp\left(\alpha_{VL}*sim\left(\boldsymbol{X}_V^p,\boldsymbol{X}_L^s\right)+\alpha_{VA}*sim\left(\boldsymbol{X}_V^p,\boldsymbol{X}_A^s\right)\right)}{\sum_{s\in S_V}exp\left(\alpha_{VL}*sim\left(\boldsymbol{X}_V^p,\boldsymbol{X}_L^s\right)+\alpha_{VA}*sim\left(\boldsymbol{X}_V^p,\boldsymbol{X}_A^s\right)\right)} \tag{36}$$

$$L_A^t=-\frac{1}{|P_A|}\sum_{p\in P_A}\log\frac{exp\left(\alpha_{AV}*sim\left(\boldsymbol{X}_A^p,\boldsymbol{X}_V^s\right)+\alpha_{AL}*sim\left(\boldsymbol{X}_A^p,\boldsymbol{X}_L^s\right)\right)}{\sum_{s\in S_A}exp\left(\alpha_{AV}*sim\left(\boldsymbol{X}_A^p,\boldsymbol{X}_V^s\right)+\alpha_{AL}*sim\left(\boldsymbol{X}_A^p,\boldsymbol{X}_L^s\right)\right)} \tag{37}$$

最终，CSD损失函数定义为所有模态对比损失的加权平均，以确保教师和学生模型之间的对齐和信息融合，具体计算过程如下所示：

$$L_{CSD}=\frac{1}{6}\left(L_L^t+L_V^t+L_A^t+L_L^s+L_V^s+L_A^s\right) \tag{38}$$

3.10 基于相似性的表征蒸馏(SRD)

教师模型基于完整模态的数据集进行训练，而学生模型可能会面临某些模态缺失的情况。现有工作[51]仅通过输出层的软标签(Soft Labels)对齐教师与学生模型，忽略了特征层级的分布差异。尤其在模态部分缺失的场景下，学生模型特征易偏离教师模型的高质量表示，导致知识传递效率下降。为此，本文提出了基于相似性的表征蒸馏(SRD)，以确保学生模型在存在模态缺失的情况下能够学习到与教师模型相一致的表征。SRD将学生模型的融合特征与教师模型的融合特征对齐，以便最大限度地保留教师模型中相对更为丰富的多模态情感交互信息，该过程可以通过公式(39)所示的相似性损失实现。

$$L_{SRD}=\alpha_{st}\left(1-\frac{\boldsymbol{W}_s\boldsymbol{H}^t\odot\boldsymbol{W}_t\boldsymbol{H}^s}{\left\|\boldsymbol{W}_s\boldsymbol{H}^t\right\|\left\|\boldsymbol{W}_t\boldsymbol{H}^s\right\|}\right) \tag{39}$$

其中，$\boldsymbol{H}^t$和$\boldsymbol{H}^s$分别表示教师和学生模型的融合特征，$\boldsymbol{W}_t$和$\boldsymbol{W}_s$分别表示可学习的教师和学生模型的特征投影矩阵，$\odot$表示逐元素乘法，$\|\ \|$表示L2正则化。通过这种训练策略，模型可以确保学生和教师之间的表示尽可能对齐，从而提高学生模型的表现。

3.11 训练目标

3.11.1 正则化损失

为了防止模型过拟合，并鼓励其学习到更加稳健的表示，本文在优化目标中引入正则化损失L_{reg}，旨在对与注意力机制相关的权重矩阵进行约束，即，对查询(Query)、键(Key)、值(Value)的投影矩阵施加惩罚。这一正则化策略能够有效抑制权重矩阵过大，从而限制模型的复杂度，同时提高模型对未见数据的泛化能力。正则化损失定义如下：

$$L_{reg}=\lambda_Q^m\left\|\boldsymbol{W}_Q^m\right\|^2+\lambda_K^m\left\|\boldsymbol{W}_K^m\right\|^2+\lambda_V^m\left\|\boldsymbol{W}_V^m\right\|^2 \tag{40}$$

其中，$\boldsymbol{W}_Q^m$、$\boldsymbol{W}_K^m$和$\boldsymbol{W}_V^m$分别表示注意力层中查询、键和值的权重矩阵，而λ_Q^m、λ_K^m和λ_V^m则是相应的正则化系数，用于控制各项正则化的相对强度。通过引入上述正则化项，本文能够在训练过程中有效降低模型对训练数据的过度拟合风险，从而增强其对多样输入数据的适应能力。这种正则化方法不仅约

束了注意力机制中关键参数的增长，还为模型在复杂任务场景中的稳健表现奠定了基础。

3.11.2　任务损失

为优化学生模型在情感预测任务上的表现，指导学生模型的优化过程，本文设计了基于交叉熵损失的任务损失L_{task}，定义如下：

$$L_{task}=-\frac{1}{N}\sum_{n=1}^{N}y_n\log\hat{y}_n^s \tag{41}$$

其中，y_n表示第n个样本的真实标签，$\hat{y}_n^s$由公式(34)计算得出，表示学生模型预测的该样本属于正确类别的概率。交叉熵损失通过衡量预测分布与真实分布之间的差异，驱动模型朝着更准确的方向进行优化。

3.11.3　综合损失函数

为确保模型在处理多模态情感分析任务时具备强大的适应性和鲁棒性，本文定义了AUMDF框架的综合损失函数，该损失函数综合了以下四个关键部分：

(1) 正则化损失L_{reg}：防止模型过拟合，增强其对未见数据的泛化能力。

(2) 对比损失L_{CSD}：如3.9节所述，用于对齐教师模型与学生模型的表示。

(3) 相似性损失L_{SRD}：如3.10节所述，旨在通过模态对齐增强模型的鲁棒性。

(4) 任务损失L_{task}：优化情感预测的最终目标。

基于以上四种损失，得到模型整体的综合损失函数L如公式(42)所示：

$$L=\lambda_1L_{reg}+\lambda_2L_{CSD}+\lambda_3L_{SRD}+L_{task} \tag{42}$$

其中，λ_1、λ_2和λ_3是损失项的权重系数，用于控制不同损失项在优化目标中的相对重要性。通过调整这些系数，模型能够根据任务需求动态调整学习重点，从而显著提升多模态数据处理的准确性和稳健性。

4　实验与结果分析

4.1　数据集与评估指标

4.1.1　数据集

本研究选用三个主流的多模态情感分析数据集：CMU-MOSI[59]、IEMOCAP[60]和CMU-MOSEI[61]，全面评估提出的AUMDF模型在不同场景下的性能表现。

CMU-MOSI数据集包含2199个视频片段，这些片段为89名独立发言者在YouTube电影评论中的观点表达。参与者中有41名女性和48名男性，其观点涵盖了从-3(非常负面)到+3(极其正面)的情感强度。该数据集被广泛用于单任务情感强度预测和情感分类任务，提供了良好的基准评估环境。

CMU-MOSEI数据集为Zadeh等人于2018年基于CMU-MOSI数据集进行的改进。相较于其前身，CMU-MOSEI在样本规模上实现了显著扩展，该数据集包含了源自5000个不同视频的22 856个视频片段，并且涵盖了更加多元化的表达者与主题。

IEMOCAP数据集主要用于多标签情感识别，包含302个视频片段，其中151个视频为对话记录。每个对话句子均标注了具体的情感类别，如快乐、悲伤、愤怒、惊讶和恐惧等，共计十余种情感标签。借鉴文献[62]中的方法，本文识别快乐、悲伤、愤怒和中立四种基础情感，以便与现有方法进行统一的对比评估。

4.1.2　评价指标

为了更精确地衡量模型的性能表现，本文采用了多种评价指标进行全面验证。与文献[63]中的工作相似，在回归任务中，本文使用平均绝对误差(MAE)衡量模型预测情感强度与实际标签之间的偏差。在分类任务中，采用准确率(Accuracy)和F1分数(F1-Score)作为衡量标准，分别用于二分类任务(正面与负面情感)和7级分类任务(如将1.8四舍五入归类为第2类情感强度)。

4.2　特征提取与实验参数设置

为了保证模型的可复现性，本文详细描述了特征提取过程、数据对齐方法、数据集划分以及实验参数设置。

4.2.1　特征提取

在三个数据集上，本文使用相同的特征提取策略，分别针对文本模态、音频模态和视觉模态进行特征提取，以确保多模态特征的一致性和可比性。

(1)文本特征提取：本文将由视频转录的文本转换为预训练的GloVe词向量[64]，将每个单词映射为300维的向量表示，这种方法能够有效捕获语言中的语义信息并提升文本特征表示的质量。

(2)音频特征提取：使用COVAREP[65]提取音频的低层特征，共计74维，包括Mel倒谱系数、基频、声调/非声调片段、归一化振幅商、声门源参数、谐波模型等。这些特征有助于捕捉音频中的情感变化和语调特征。

(3)视觉特征提取：视觉特征由MA-Net[66]提取，该网络因其在面部表情识别中的突出表现而被

广泛应用。本文首先使用MTCNN模型进行面部检测，随后利用预训练的MA-Net模型提取每帧的1024维作为视频特征。

为实现多模态的词级对齐，本文利用P2FA工具[67]对音频和视频流进行时间戳对齐，并在这些同步间隔内对音频和视频特征进行平均。对于CMU-MOSI数据集与CMU-MOSEI数据集，将每个模态的序列长度设定为50，而在IEMOCAP数据集上，所有模态的序列长度均设为20。

本文按照固定比例将数据集划分为训练集、验证集和测试集，具体分割统计见表1。

表1 数据集划分情况

数据集	CMU-MOSI	CMU-MOSEI	IEMOCAP
训练集	1284	16 326	2717
验证集	229	1871	798
测试集	686	4659	938

4.2.2 实验设置

所有实验均基于PyTorch框架完成，运行环境包括NVIDIA Tesla V100 GPU，PyTorch版本为1.8.2。为了确保结果的公平性和可比性，本文重新实现了几种现有的先进方法，并基于公开的代码库进行了模型集成和比较。

本文参考文献[68]提出的模型进行超参数配置，模型训练过程中主要参数的设置如表2所示。模型的学习率设为0.001，隐藏层维度设为300，优化器采用Adam，总训练轮次为20。此外，为了控制模型的过拟合风险，本文设定了dropout率为0.8，并引入早停策略(early stop)，以监控验证集的性能并提前停止训练。

表2 实验超参数设置

超参数	符号定义	数值设置
训练轮次	b	20
Dropout率	d	0.8
隐藏层大小	h	300
学习率	lr	0.001
丢失率	p	[0.1−0.7]
最大文本序列长度	m_l	25
最大音频序列长度	m_a	150
最大视觉序列长度	m_v	100
损失函数权重	$\lambda_1,\lambda_2,\lambda_3$	0.1
早停轮次	es	20

4.3 对比算法

为了验证AUMDF的有效性，本文与多组可复现的代表性先进方法进行了详细对比，下面对这些方法进行简要介绍。

MISA[69]：Modality-Invariant and-Specific Representations(模态不变和模态特定表示)模型为每个模态分别分配两个不同的表示空间，一个用于捕获模态间的共性并逐步缩小模态间的差异，另一个用于专注每个模态的独特特性。最终，这两个表示空间被融合以实现更精确的情感分析。

Self_MM[63]：Self-Supervised Multi-Modal Learning Framework(自监督多模态学习框架)专为多模态和单模态任务的多任务学习而设计。该方法引入了一个基于自监督学习策略的标签生成模块以便在单模态任务中生成标签。此外，为了在训练过程中平衡各子任务的学习进度，设计了一种动态权重调整机制，以提升不同任务的协同学习效果。

CMJRT[70]：Cross-Modal Joint Representation Translator(跨模态联合表示翻译器)首先学习每对模态之间的关联关系，并利用跨模态Transformer探索模态之间的互补性，从而增强多模态信息的综合利用。

EMT[40]：Efficient Multimodal Transformer(高效多模态Transformer)通过获取每种模态的全局特征，充分挖掘单一模态的局部特征交互。为了增强模型在模态缺失场景下的鲁棒性，该模型引入了双层特征恢复机制(DLFR)，以便从有限的数据中获取更丰富的语义信息。

CubeMLP[71]：Cube Multilayer Perceptron(立方多层感知机)利用三个独立的MLP单元分别对三种模态的特征进行有效信息的学习与融合，从而通过聚合的多模态特征实现更准确的任务预测。这种方法充分利用了简单高效的神经网络结构来降低计算复杂度。

MMIN[72]：Missing Modality Imagination Network(缺失模态想象网络)是一个连贯的多模态情感检测框架，该模型利用模态重构注意力机制(CRA)和递归一致性学习方法来推断缺失的数据，使其在多模态数据不完整的情况下依然能保持良好的性能。

IF-MMIN[73]：Invariant Features-Missing Modality Imagination Network(不变特征缺失模态想象网络)利用现有模态预测缺失模态的固有特征，能够在模态缺失的情况下准确推测并推理出丢失模态的数据特征进行情感分析，从而实现更高的预测性能。

4.4　实验结果分析

4.4.1　不同模态组合下的实验结果

为进一步研究AUMDF模型在多模态缺失条件下的鲁棒性，本文在CMU-MOSI、CMU-MOSEI和IEMOCAP数据集上进行了多组实验，分别测试了模型在不同模态组合下的表现。实验结果如表3至表5所示，其中，$\{l\}$表示只包含文本模态，$\{l,a,v\}$表示包含文本、音频与视觉模态。

从在三个数据集上的实验结果可以看出，尽管AUMDF的个别指标表现略低于其他对比的基线模型，但大部分指标相比其他基线模型均表现出优越性能。在部分模态缺失的情况下，AUMDF依然能够取得显著的性能提升，这表明其具备良好的抗干扰能力和模态自适应性。

表3　在CMU-MOSI数据集上的实验结果

	不同模态组合															
	$\{l\}$		$\{a\}$		$\{v\}$		$\{l,a\}$		$\{l,v\}$		$\{a,v\}$		$\{l,a,v\}$			
模型	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	Avg_MAE	Avg_F1
MISA	0.796	66.93	0.756	42.82	0.741	40.71	0.795	65.49	0.691	67.54	0.817	48.36	0.798	80.21	0.771	58.87
Self-MM	0.787	67.8	0.758	40.95	0.747	38.52	0.748	69.81	0.739	74.97	0.845	47.12	0.809	84.09	0.776	60.47
CMJRT	0.803	68.79	0.785	43.23	0.763	42.62	0.695	70.15	0.729	68.54	0.823	50.71	0.887	82.83	0.784	60.98
EMT	0.823	75.12	0.768	59.52	0.792	58.75	0.703	**77.18**	0.705	74.28	0.828	61.35	0.735	83.22	0.765	69.92
CubeMLP	0.838	64.15	0.801	38.91	0.731	43.24	0.725	63.76	0.729	65.12	0.839	47.92	0.782	83.27	0.778	58.05
MMIN	0.822	55.73	0.833	46.47	0.807	44.71	0.729	61.88	0.778	64.11	0.805	67.4	0.765	81.85	0.791	60.31
IF-MMIN	0.849	57.86	0.823	**66.21**	0.816	50.62	0.732	58.91	0.732	68.98	0.796	65.39	0.787	82.02	0.790	64.28
AUMDF	**0.762**	**81.31**	**0.754**	64.53	**0.735**	**60.84**	**0.702**	73.68	**0.686**	**79.41**	**0.789**	**73.84**	**0.729**	**84.37**	**0.737**	**73.99**

表4　在CMU-MOSEI数据集上的实验结果

	不同模态组合															
	$\{l\}$		$\{a\}$		$\{v\}$		$\{l,a\}$		$\{l,v\}$		$\{a,v\}$		$\{l,a,v\}$			
模型	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	MAE	F1	Avg_MAE	Avg_F1
MISA	0.751	68.94	0.733	43.87	0.727	41.76	0.779	69.48	0.682	72.55	0.806	48.91	0.754	82.7	0.747	61.17
Self-MM	0.764	71.53	0.742	43.57	0.739	37.61	0.724	75.91	0.728	74.62	0.827	49.52	0.731	83.05	0.751	62.26
CMJRT	0.789	70.12	0.786	42.39	0.758	38.47	0.687	74.54	0.716	72.63	0.817	50.21	0.809	81.76	0.766	61.45
EMT	0.802	72.34	0.697	44.21	0.781	39.14	0.695	74.23	0.694	73.61	0.841	62.95	0.721	82.73	0.747	64.17
CubeMLP	0.785	67.52	**0.783**	39.54	0.712	32.58	0.703	71.69	0.701	70.06	0.826	48.54	0.783	83.17	0.756	59.01
MMIN	0.816	68.23	0.751	41.45	0.785	33.92	0.712	71.25	0.754	70.71	0.799	48.91	0.741	82.75	0.765	59.60
IF-MMIN	0.803	69.44	0.789	42.51	0.807	34.92	0.728	73.21	0.722	71.19	0.775	50.32	0.773	82.96	0.771	60.65
AUMDF	**0.742**	**72.76**	0.735	**57.57**	**0.716**	**58.41**	**0.679**	**79.78**	**0.674**	**81.06**	**0.733**	**71.54**	**0.705**	**83.29**	**0.712**	**72.06**

如表3所示，在CMU-MOSI数据集的完整模态组合（$\{l,a,v\}$）下，AUMDF取得了最低的MAE和最高的F1分数，显著优于基于自监督学习策略的Self-MM模型。这一结果表明，当所有模态可用时，AUMDF得益于基于知识蒸馏的特征训练策略加入，能够更有效地捕捉模态间的关联性，缓解了Self-MM模型中因模态交互不足导致的学习不平衡问题，从而更好地融合各模态的情感信息。

去掉音频模态（$\{l,v\}$）时，AUMDF的MAE与F1分数分别比基于Transformer的EMT模型降低了6.5%和提高了2.3%。这是由于AUMDF在音频缺失时，相较于EMT模型的Transformer，通过设计更加灵活的多模态掩码Transformer，实现了模态信息的自适应调整，在音频模态缺失的情况也能够充分捕获文本和视觉模态中的情感特征，从而在评价指标上保持领先优势。

在仅保留文本模态（$\{l\}$）的情况下，AUMDF的F1分数和MAE比基于模态不变和模态特定表示机制的MISA模型分别提高了14.4%和降低了4.3%。可见，AUMDF在文本模态单独存在的情况下仍然表现优异。这是由于AUMDF中的动态权重调整模块弥补了MISA模型中模态特定表示与模态不变表示之间的不平衡问题，展现出了较强的适应性。

表5 在IEMOCAP数据集上的实验结果

模型	情感类别	不同模态组合 {l}	{a}	{v}	{l,a}	{l,v}	{a,v}	{l,a,v}	Avg_F1
MISA	快乐	67.3	52.7	50.8	73.2	70.3	61.3	89.2	66.4
	悲伤	68.4	52.3	53.1	69.5	68.7	61.1	88.4	65.9
	愤怒	66.9	53.5	52.7	67.7	69.5	56.6	86.1	64.7
	中立	56.4	48.4	50.9	58.1	56.5	52.8	69.2	56.0
Self-MM	快乐	66.9	52.2	50.1	69.9	68.3	56.3	**90.8**	64.9
	悲伤	68.7	51.9	54.8	71.3	69.5	57.5	86.7	65.8
	愤怒	65.4	53.0	51.9	69.5	67.7	56.6	85.4	64.2
	中立	55.8	48.2	50.4	58.1	56.5	52.8	70.7	56.1
CMJRT	快乐	69.2	55.3	52.5	79.6	77.2	65.8	83.1	69.0
	悲伤	65.4	55.2	53.9	78.9	73.1	68.4	82.8	68.2
	愤怒	65.1	54.6	51.3	81.6	80.3	59.8	84.6	68.2
	中立	54.3	51.5	49.2	65.7	62.4	54.9	67.1	57.9
EMT	快乐	77.2	63.8	61.3	81.6	80.2	66.5	85.5	73.7
	悲伤	76.3	64.2	60.9	82.4	81.5	64.8	84.0	73.4
	愤怒	77.6	61.8	58.2	83.9	81.7	**68.2**	85.1	73.8
	中立	60.7	51.5	50.4	65.2	62.4	56.8	67.1	59.2
CubeMLP	快乐	68.9	54.3	51.4	72.1	69.8	60.6	89.0	66.6
	悲伤	65.3	54.8	53.2	70.3	68.7	58.1	88.5	65.6
	愤怒	65.8	53.1	50.4	69.5	69.0	54.8	86.1	64.1
	中立	53.5	50.8	48.7	57.3	54.5	51.8	**71.8**	55.5
MMIN	快乐	80.6	66.5	64.2	83.1	81.8	67.2	90.3	76.2
	悲伤	79.1	65.3	62.8	82.4	79.6	70.4	85.4	75.0
	愤怒	80.3	**67.4**	61.6	83.0	82.3	59.7	84.9	74.2
	中立	61.0	50.7	49.5	62.2	52.6	55.2	67.2	56.9
IF-MMIN	快乐	82.4	67.7	66.9	83.5	**82.6**	69.8	87.3	77.2
	悲伤	81.1	69.0	66.3	**84.1**	81.8	70.3	86.9	77.1
	愤怒	81.9	67.1	65.5	82.5	81.6	68.1	85.2	76.0
	中立	61.4	52.6	43.1	64.9	62.7	57.2	71.5	59.1
AUMDF	快乐	**82.7**	**69.7**	**68.1**	**84.2**	82.1	**70.1**	87.6	**77.8**
	悲伤	**82.4**	**71.4**	**67.5**	83.3	**82.4**	**72.4**	**88.7**	**78.3**
	愤怒	**82.3**	67.1	**66.3**	**83.6**	**82.9**	67.3	**86.2**	**76.5**
	中立	**63.5**	**54.3**	**52.4**	**68.3**	**64.5**	**57.8**	71.2	**61.7**

如表4所示，在CMU-MOSEI数据集的完整模态组合（$\{l,a,v\}$）下，其MAE值取得最低，然而F1分数略低于Self-MM。这一结果可能由于知识蒸馏策略导致了部分高频情感语义（特别是极端情感类别）的过度平滑，进而使得F1分数略有下降。尽管如此，AUMDF在平均MAE与平均F1分数上仍然表现最佳，这一表现得益于多模态掩码Transformer对跨模态间隐式关联的强大学习能力，以及知识蒸馏策略对原始特征进行的精细化再构建。

如表5所示，在IEMOCAP数据集的完整模态组合（$\{l,a,v\}$）下，AUMDF在“快乐”、“悲伤”、“愤怒”以及“中立”四种情感类别上的F1分数相较于基于现有特征的模态补充机制的IF-MMIN模型分别提高了6.1%、5.5%、1.2%和18.6%。这是由于AUMDF的动态调整机制在不引入冗余噪声的同时，实现了信息的深度融合，因此AUMDF的指标优于IF-MMIN模型。

当去掉视觉模态（$\{l,a\}$）时，AUMDF在中立类别上的F1分数远高于基于跨模态Transformer的CMJRT模型。这是由于AUMDF中的注意力机制能够在模态缺失时优先捕获剩余模态中的关键特征，弥补了缺失模态带来的信息缺口。

为进一步验证AUMDF在不同模态丢失率下的性能，我们在三个数据集上进行了丢失率从0.2到0.8的实验。实验结果如图4至图6所示。当模态内丢失率从0.2上升至0.8时，AUMDF的F1分数始终优于其他模型。例如，在丢失率为0.4时，AUMDF的F1分数相较Self-MM和CMJRT分别提升了6.8%和11.4%。这一表现得益于AUMDF中的基于知识蒸馏的特征训练策略，通过对比学习有效地将教师模型的完整模态信息传递给学生模型，从而在缺失模态的情况下仍能保留关键信息。

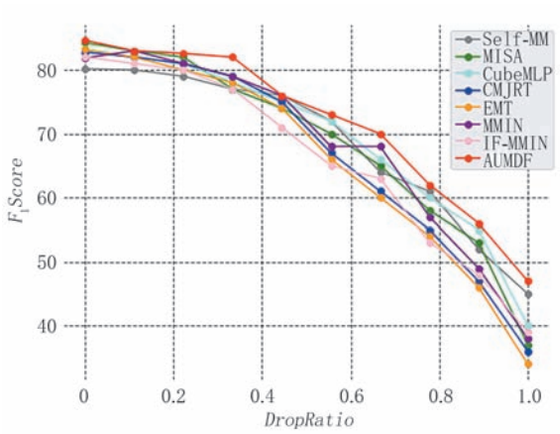

图4 不同丢失率设置下在CMU-MOSI数据集上的实验结果

上述实验结果进一步验证了AUMDF在多模态情感分析任务中的有效性。未来可以考虑结合更先进的知识蒸馏策略，以进一步提升模型在高丢失率条件下的性能。

4.4.2 消融实验

为了全面验证本文提出的AUMDF中各个模块与损失函数对多模态情绪识别性能的具体影响，并进一步探讨模型在处理模态缺失和多模态信息融合中的有效性，我们进行了消融实验。实验在模态

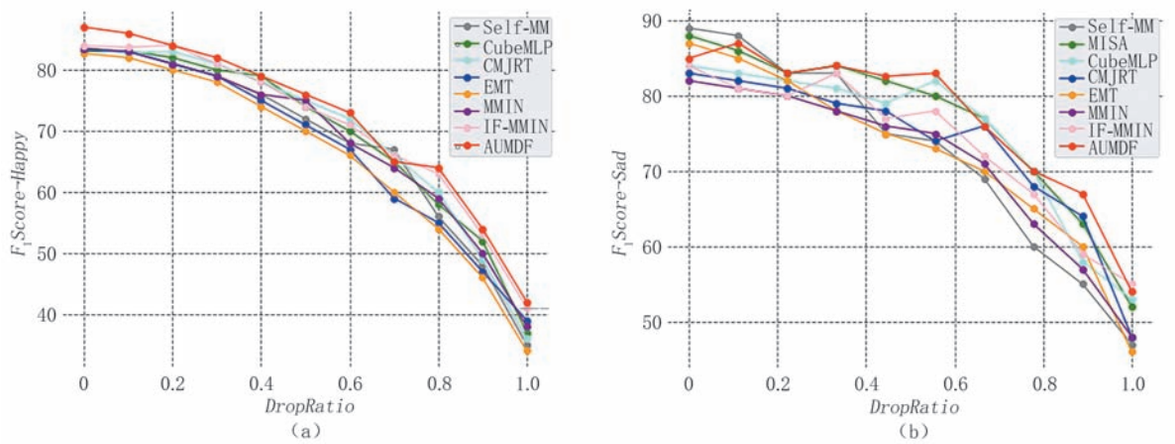

图5　不同丢失率设置下在IEMOCAP数据集上的实验结果

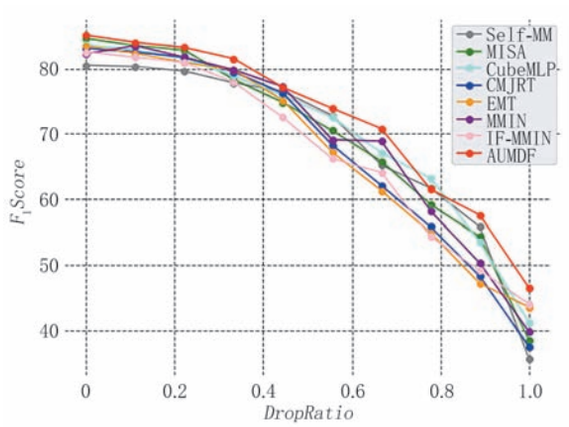

图6　不同丢失率设置下在CMU-MOSEI数据集上的实验结果

丢失率 p=0.2的设定下进行，表6和表7分别展示了基于动态权重调整模块(DWAM)、对比样本蒸馏(CSD)与基于相似性的表征蒸馏(SRD)的消融实验详细结果。出于篇幅限制，在CMU-MOSEI数据集与在IEMOCAP数据集上的实验趋势与本文一致，此处不再展示。

表6　丢失率p=0.2时在CMU-MOSI数据集上的模块消融实验结果(√表示包含此模块)

DWAM	CSD	SRD	MAE	F1
			0.820	82.36
	√	√	0.782	84.15
√		√	0.798	83.82
√	√		0.764	84.09
√	√	√	**0.758**	**84.57**

模块消融实验旨在评估模型中的各模块对整体性能的贡献。在每次实验中，依次移除特定模块并观察其对情绪识别任务的影响，表5展示了各模块

表7　在CMU-MOSI数据集上的损失函数消融实验结果

损失函数组合	MAE	F1
L_{task}	0.896	76.39
$L_{task}+L_{reg}$	0.871	77.13
$L_{task}+L_{reg}+L_{CSD}$	0.849	79.84
$L_{task}+L_{reg}+L_{SRD}$	0.837	80.06
$L_{reg}+L_{CSD}+L_{SRD}$	0.793	81.43
$L_{reg}+L_{CSD}+L_{SRD}+L_{task}$	**0.684**	**84.55**

消融的结果。

从模块消融实验的结果中可以看出以下几点关键结论：

(1) 移除动态权重调整模块(DWAM)后，模型的MAE显著增加8.2%，F1得分则减少2.6%。这是由于DWAM模块的主要作用是动态调整多模态特征之间的权重，使得重要特征在不同情境下能够获得更高的权重。而移除DWAM后，模型失去了这种动态调节能力，导致在融合过程中难以准确区分关键情绪信息，尤其在处理模态间信息不均衡时，性能出现显著下降。

(2) 移除对比样本蒸馏机制(CSD)后，模型的MAE增加3.2%，F1得分减少0.5%。与DWAM相比，CSD模块的影响相对较小，但依然显著。这是由于CSD的作用在于确保不同模态特征能够在语义空间中保持一致性，从而更好地捕获模态间的潜在交互关系。其移除导致模型在模态融合过程中信息对齐能力减弱，最终对情绪识别的细粒度精度产生了负面影响。

(3) 移除基于相似性的表征蒸馏机制(SRD)后，模型的MAE增加5.3%，F1得分减少0.9%。这是由于SRD模块的核心功能是通过类别原型的引导，提升模型在复杂情绪类别上的区分能力。其

缺失导致模型在应对噪声数据时的鲁棒性显著降低，尤其在类别边界较为模糊的情绪样本中，分类精度下降明显。

上述三组消融实验结果进一步证明了AUMDF模型各模块的合理性与有效性，并可以得出如下结论：动态权重调整模块对情绪信息的动态调节能力最为显著，对比样本蒸馏在增强模型一致性方面具有积极作用，而基于相似性的表征蒸馏则通过增强类别区分能力进一步提升了模型的整体性能。

为进一步验证各损失函数对模型性能的具体贡献，本文对AUMDF进行了损失函数消融实验。表6-7展示了AUMDF在CMU-MOSI数据集上的不同损失函数组合下的实验结果。出于篇幅限制，在CMU-MOSEI数据集与在IEMOCAP数据集上的实验趋势与本文一致，此处不再展示。

（1）仅使用任务损失（L_{task}）

当模型仅使用L_{task}时，模型性能显著下降。这是因为单一任务损失仅能优化情绪分类任务本身，而无法充分考虑模态间的信息互补与对齐，从而导致情绪特征融合不足，模型在应对模态缺失情境时性能下降明显。

（2）加入正则化损失（$L_{task}+L_{reg}$）

在任务损失基础上增加正则化损失L_{reg}后，模型的MAE降至0.871，F1得分略有提升。这说明正则化损失在约束模型参数、减少过拟合方面具有显著作用，有助于增强模型的泛化能力。

（3）加入相似性损失（$L_{task}+L_{reg}+L_{CSD}$）

当进一步引入L_{CSD}时，模型的MAE进一步降低，F1得分则有所提升。这表明，相似性损失有效确保了各模态特征在表示上的协调性，从而显著提升了情绪分类的精度与稳定性。

（4）任务损失（$L_{reg}+L_{CSD}+L_{SRD}+L_{task}$）

在使用完整的任务损失的情况下，模型达到了最佳性能，相比于仅使用L_{task}时，MAE减少了23.7%，F1得分提升了8.16%。此结果充分证明了本文设计的任务损失在提升模型鲁棒性与多模态情绪识别性能方面的有效性。

4.5 可视化分析

为了更直观地展示AUMDF模型在不同模态丢失情况下的表现，本文从IEMOCAP测试集中随机选取了四类情绪样本（快乐、愤怒、中性、悲伤）进行可视化分析，采用t-SNE算法进行降维，将多模态融合特征从高维空间映射到二维空间，使得样本聚类和类别间分离更加清晰，结果如图7所示。

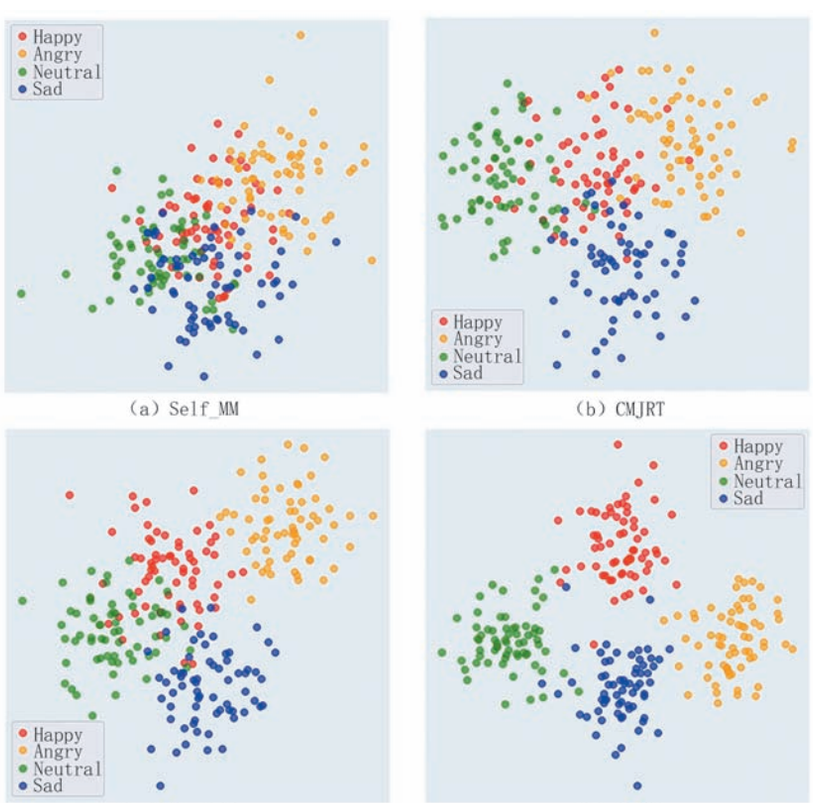

图7 在IEMOCAP测试集上的可视化结果

如图7(a)所示，Self-MM模型在模态丢失情况下各情绪类别间表示高度重叠，说明其对模态信息缺失的鲁棒性较差，难以有效区分不同情绪类别。

如图7(b)和7(c)所示，这两种模型在情绪区分上有所改善，但不同类别样本之间仍然存在明显重叠，说明其在特征融合方面仍然不够充分。

如图7(d)所示，AUMDF模型在模态丢失情况下依然能保持各情绪类别之间的清晰边界，同类别样本聚集紧密，跨类别样本分布显著。这证明了AUMDF框架在应对模态缺失、多模态融合任务中的优越性和鲁棒性。

5　结　论

为解决多模态情感分析中模态缺失带来的问题，本文提出了一种基于知识蒸馏的多模态情感分析模型AUMDF。该模型旨在通过设计有效的跨模态特征融合模块与特征训练策略，克服不同模态信息不完整或缺失所导致的性能下降问题。AUMDF对原始特征增强优化后，基于重要性对各个模态进行自适应权重分配，挖掘模态间的相关性，再通过多模态掩码Transformer结构继续捕捉模态间的深层细粒度交互信息，最后设计蒸馏学习机制实现教师与学生模型间的表征对齐，从而在模态缺失环境中提升情感分析的准确性。在IEMOCAP、CMU-MOSI和CMU-MOSEI数据集上的实验结果表明，AUMDF在模态缺失场景中的性能显著优于现有方法，尤其在高丢失率条件下展现了优秀的适应能力和泛化性能。通过动态融合和知识蒸馏，AUMDF为多模态情感分析提供了一种高效且鲁棒的解决方案，为未来该领域研究提供了新思路和参考框架。

参　考　文　献

[1] Xue X, Zhang C, Niu Z, et al. Multi-level attention map network for multimodal sentiment analysis. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(5): 5105-5118

[2] Liu Z, Zhou B, Chu D, et al. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 2024, 101: 101973

[3] Rozgić V, Ananthakrishnan S, Saleem S, et. al. Ensemble of SVM trees for multimodal emotion recognition//Proceedings of the 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference. Hollywood, California, USA, 2012: 1-4

[4] Cummins N, Amiriparian S, Ottl S, et al. Multimodal bag-of-words for cross-domains sentiment analysis//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada, 2018: 4954-4958

[5] Wang M, Cao D, Li L, et al. Microblog sentiment analysis based on cross-media bag-of-words model//Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMCS′14). New York, USA, 2014: 76-80

[6] Yu J, Jiang J, Xia R. Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 429-439

[7] Yu J, Jiang, J. Adapting BERT for target-oriented multimodal sentiment classification//Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI′19). Macao, China. 2019: 5408-5414

[8] Wang J, Liu Z, Sheng V, et al. SaliencyBERT: recurrent attention network for target-oriented multimodal sentiment classification//Proceedings of the 4th Chinese Conference on Pattern Recognition and Computer Vision (PRCV) 2021. Beijing, China, 2021: 3-15

[9] Pan T, Ye Y, Cai H, et al. Multimodal physiological signals fusion for online emotion recognition//Proceedings of the 31st ACM International Conference on Multimedia (MM′23). New York, USA, 2023: 5879-5888

[10] Li B, Fei H, Liao L, et al. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition//Proceedings of the 31st ACM International Conference on Multimedia (MM′23). New York, USA, 2023: 5923-5934

[11] Chen F, Luo Z, Xu Y, et al. Complementary fusion of multi features and multi-modalities in sentiment analysis//Proceedings of the CEUR Workshop. Copenhagen, Denmark, 2020: 82-89

[12] Yang J, Xiao Y, Du X. Multi-grained fusion network with self-distillation for aspect-based multimodal sentiment analysis. Knowledge-Based Systems, 2024, 293(C), 101-119

[13] Acheampong, F, Nunoo-Mensah, H, Chen W. Transformer models for text-based emotion detection: A review of BERT-based approaches. Artificial Intelligence Review, 2021, 54(8): 5789-5829

[14] Arumugam B, Bhattacharjee S, Yuan J. Multimodal attentive learning for real-time explainable emotion recognition in conversations//Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS). Austin, USA, 2022: 1210-1214

[15] Castro S, Hazarika D, Pérez-Rosas V, et al. Towards multimodal sarcasm detection//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4619-4629

[16] Cai Y, Cai H, Wan X. Multi-modal sarcasm detection in Twitter with hierarchical fusion model//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL′19). Florence, Italy, 2019: 2506-2515

[17] Huddar M, Sannakki S, Rajpurohit V. Attention-based

multimodal contextual fusion for sentiment and emotion classification using bidirectional LSTM. Multimedia Tools and Applications, 2021, 80(9): 13059-13076

[18] Zhang Q, Shi L, Liu P, et al. ICDN: Integrating consistency and difference networks by transformer for multimodal sentiment analysis. Applied Intelligence, 2022, 24(6):29-37

[19] Morency L, Mihalcea R, Doshi P. Towards multimodal sentiment analysis: harvesting opinions from the web//Proceedings of the 13th International Conference on Multimodal Interfaces (ICMI 2011). Alicante, Spain, 2011: 169-176

[20] Pérez-Rosas V, Mihalcea R, Morency L. Utterance-level multimodal sentiment analysis//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013). Sofia, Bulgaria, 2013: 973-982

[21] Zadeh A, Liang P P, Poria S, et al. Multi-attention recurrent network for human communication comprehension//Proceedings of the 32th AAAI Conference on Artificial Intelligence. New Orleans , USA , 2018: 5642-5649

[22] Shutova E, Kiela D, Maillard J. Black holes and white rabbits: metaphor identification with visual features//Proceedings of the NAACL HLT 2016. San Diego, USA, 2016: 160-170

[23] Morvant E, Habrard A, Ayache S. Majority vote of diverse classifiers for late fusion//Proceedings of the Structural, Syntactic, and Statistical Pattern Recognition Joint IAPR International Workshop (S+SSPR 2014). Joensuu, Finland, 2014: 153-162

[24] Evangelopoulos G, Zlatintsi A, Potamianos A, et al. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Transactions on Multimedia, 2013, 15(7): 1553-1568

[25] Luo Yuan-Yi, Wu Rui, Liu Jiafeng, et al. A multimodal sentiment analysis approach for emotion semantic inconsistencies 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/OL].计算机研究与发展, 2025, 62(2): 374-382)

[26] Arevalo J, Solorio T, Gómez M, et al. Gated multimodal units for information fusion//Proceedings of the International Conference on Learning Representations (ICLR). Toulon, France, 2017: 1-17

[27] Yang H, Gao X, Wu J, et al. Self-adaptive context and modal-interaction modeling for multimodal emotion recognition//Proceedings of the Findings of the Association for Computational Linguistics (ACL 2023). Toronto, Canada, 2023: 6267-6281

[28] Zong Linlin, Zhou Jiahui, Xie Qiujie, et al. Multimodal emotion recognition based on hypergraph. Chinese Journal of Computers, 2023, 46(12):2520-2534 (in Chinese)
(宗林林,周佳慧,谢秋婕,等.基于超图的多模态情绪识别.计算机学报, 2023, 46(12):2520-2534)

[29] Zhang Z, Lan C, Zeng W, et al. Relation-aware global attention for person re-identification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020). Seattle, USA, 2020: 3183-3192

[30] Lian Z, Liu B, Tao J. CTNet: conversational transformer network for emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 985-1000

[31] Mai S, Zeng Y, Zheng S, et al. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. IEEE Transactions on Affective Computing, 2023, 14(3): 2276-2289

[32] Yadav A, Vishwakarma D. A deep multi-level attentive network for multimodal sentiment analysis. ACM Transactions on Multimedia Computing, Communications, and Applications, 2023, 19(1):1-19

[33] Dai R, Tan Y, Mo L, et al. MuAP: multi-step Adaptive Prompt Learning for Vision-Language Model with Missing Modality. ArXiv Preprint ArXiv: 2409.04693, 2024

[34] Zhang H, Wang W, Yu T. Towards robust multimodal sentiment analysis with incomplete data. ArXiv Preprint ArXiv: 2409.20012, 2024

[35] Ma M, et al. SMIL: Multimodal learning with severely missing modality//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2021: 2302-2310

[36] Yuan Z, Fang J, Xu H, et al. Multimodal consistency-based teacher for semi-supervised multimodal sentiment analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2024, 32: 3669-3683

[37] Li M, Yang D, Liu Y, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. ArXiv Preprint ArXiv: 2411.02793, 2024

[38] Pham H, Liang P, Manzini T, et al. Found in translation: learning robust joint representations by cyclic translations between modalities//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, Hawaii, USA, 2019: 6892-6899

[39] Yuan Z, Liu Y, Xu H, et al. Noise imitation based adversarial training for robust multimodal sentiment analysis. IEEE Transactions on Multimedia, 2024, 26:529-539

[40] Sun L, Lian Z, Tao L J. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. IEEE Transactions on Affective Computing, 2024, 15(1):309-325

[41] Shi P, Hu M, Nakagawa S, et al. Text-guided reconstruction network for sentiment analysis with uncertain missing modalities. IEEE Transactions on Affective Computing, 2025, 1:1-15

[42] Lian Z, Chen L, Sun L, et al. GCNet: graph completion network for incomplete multimodal learning in Conversation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(7):8419-8432

[43] Wu Y, Lin Z, Zhao Y, et al. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Bangkok, Thailand, 2021: 4730-4738

[44] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. ArXiv Preprint ArXiv:1503.02531, 2015

[45] Wei P, Yang J, Xiao Y. Hierarchical cross-modal interaction

and fusion network enhanced with self-distillation for emotion recognition in conversations. Electronics, 2024, 13: 2645

[46] Zhang Y, Liu F, Zhuang X, et al. Prototype-based sample-weighted distillation unified framework adapted to missing modality sentiment analysis. Neural Networks, 2024, 177: 106397

[47] Kumar S, Banerjee B, Chaudhuri S. Online sensor hallucination via knowledge distillation for multimodal image classification. ArXiv Preprint ArXiv:1908.10559, 2019

[48] Weng Y, et al. Enhancing multimodal sentiment analysis for missing modality through self-distillation and unified modality cross-attention. ArXiv Preprint ArXiv:2410.15029, 2024

[49] Wang H, Ma C, Zhang J, et al. Learnable cross modal knowledge distillation for multi-modal learning with missing modality//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Vancouver, Canada, 2023: 216-226

[50] Sun Y, Liu Z, Sheng Q, et al. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. Information Fusion, 2024, 110: 102454

[51] Wei S, Luo Y, Luo C. MMANet: margin-aware distillation and modality-aware regularization for incomplete multimodal learning//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada, 2023: 20039-20049

[52] Deng Y, Bian J, Wu S, et al. Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition. Information Fusion, 2025, 114: 102711

[53] Yuan, Z, Li W, Xu H, et al. Transformer-based feature reconstruction network for robust multimodal sentiment analysis//Proceedings of the ACM Multimedia Conference (MM′21). Chengdu, China, 2021: 4400-4407

[54] Huang C, Zhang J, Wu X, et al. TeFNA: text-centered fusion network with cross-modal attention for multimodal sentiment analysis. Knowledge-Based Systems, 2023, 269:110502

[55] Mai S, Xing S, Hu H. Analyzing multimodal sentiment via acoustic-and visual-LSTM with channel-aware temporal convolution network. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1424-1437

[56] Tsai Y H H, Bai S, Liang P P, et al.Multimodal transformer for unaligned multimodal language sequences//Proceedings of the conference. Association for Computational Linguistics (ACL). Meeting. Florence, Italy, 2019, 6558-6569

[57] Hu M, Maillard M, Zhang Y, et al. Knowledge distillation from multi-modal to mono-modal segmentation networks//Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI 2020), Lima, Peru, 2020: 772-781

[58] Pan Y, Jiang J, Jiang K, et al. Disentangled-multimodal privileged knowledge distillation for depression recognition with incomplete multimodal data//Proceedings of the 32nd ACM International Conference on Multimedia (MM′24). Association for Computing Machinery, New York, USA, 5712-5721

[59] Zadeh A, Zellers R, Pincus E, et al. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. ArXiv Preprint ArXiv:1606.06259, 2016

[60] Busso C, Bulut M, Kazemzadeh A, et al. IEMOCAP: interactive emotional dyadic motion capture database. Language Resources and Evaluation, 2008, 42: 335-359

[61] Zadeh A, Liang PP, Poria S, Cambria E, Morency LP. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Melbourne, Australia, 2018. 2236-2246

[62] Wang Y, Shen Y, Liu Z, et al. Words can shift: dynamically adjusting word representations using nonverbal behaviors//Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). Honolulu, USA, 2019: 7216-7223

[63] Yu W, Xu H, Yuan Z. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2021, 35(12): 10790-10797

[64] Pennington J, Socher R, Manning C. GloVe: global vectors for word representation//Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, 2014: 1532-1543

[65] Degottex G, Kane J, Drugman T, et al. COVAREP—A collaborative voice analysis repository for speech technologies//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Florence, Italy, 2014: 960-964

[66] Zhao Z, Liu Q, Wang S. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. IEEE Transactions on Image Processing, 2021, 30: 6544-6556

[67] Penn Phonetics Laboratory. p2fa-vislab: A script for audio/transcript alignment. GitHub, 2013. Available from: https://github.com/ucbvislab/p2favislab/

[68] Zeng J, Liu T, Zhou J. Tag-assisted multimodal sentiment analysis under uncertain missing modalities//Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. Madrid, Spain, 2022: 1545-1554

[69] Hazarika D., Zimmermann R., & Poria, S. MISA: modality-invariant and specific representations for multimodal sentiment analysis//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA, 2020: 1122-1131

[70] Xu M, Liang F, Su X, et al. CMJRT: cross-modal joint representation transformer for multimodal sentiment analysis. IEEE Access, 2022, 10: 131671-131679

[71] Sun H, Wang H, Liu J, et al. CubeMLP: an MLP-based model for multimodal sentiment analysis and depression estimation//Proceedings of the 30th ACM International Conference on Multimedia (ACM MM). Lisbon, Portugal, 2022: 3722-3729

[72] Zhao J, Li R, Jin Q. Missing modality imagination network for emotion recognition with uncertain missing modalities//

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2021: 2608-2618

[73] Zuo H, Liu R, Zhao J, et al. Exploiting modality-invariant features for robust multimodal emotion recognition with missing modalities//Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Rhodes Island, Greece, 2023: 1-5

WANG Nan, Ph. D., professor. Her research interests include machine learning, natural language processing, multimodal learning.

WANG Qi, M. S. candidate. Her main research interest is multimodal learning.

OUYANG Dan-Tong, Ph. D., professor. Her main research interests focus on model-based diagnosis and automatic reasoning.

Background

With the rapid advancement of artificial intelligence, the ability to recognize and analyze emotions has become essential for improving human-computer interaction. Consequently, multimodal sentiment analysis (MSA) has emerged as a key area of research, leveraging multiple modalities such as text, audio, and visual data to achieve superior sentiment prediction. Compared to unimodal approaches, MSA provides the advantages of data complementarity and robustness. However, achieving effective feature fusion between modalities remains a central challenge, as the quality of feature interaction directly impacts the performance of sentiment analysis.

Existing methods for MSA primarily focus on three types of fusion techniques: neural network-based, attention-based, and graph-based approaches. Neural network-based methods, such as those using Recurrent Neural Networks (RNNs) or Convolutional Neural Networks (CNNs), have demonstrated strong representation capabilities but often suffer from high computational costs and long training times. Attention-based methods, particularly those using self-attention mechanisms, effectively capture temporal dependencies but face limitations in fully leveraging inter-modal relationships, especially in scenarios involving missing modalities. This design limits their ability to handle data from three or more modalities without introducing redundant information, leading to suboptimal feature fusion performance.

Despite significant progress, handling missing modalities remains an unresolved issue in MSA research. Real-world applications often encounter incomplete or noisy data, which can severely impact model performance. Current solutions, including imputation techniques and knowledge distillation frameworks, partially mitigate these issues but fail to fully capture the complexity of incomplete multimodal data. As a result, there is a pressing need for more robust and adaptable methods to address missing modality scenarios while maintaining effective feature fusion.

This study seeks to bridge this gap by introducing the Attention-based Uncertain Missing Modality Distillation Framework (AUMDF). By incorporating novel mechanisms such as the Dynamic Weight Adjustment Module (DWAM) and Multimodal Masked Transformer (MMT), AUMDF dynamically adjusts to missing modalities while enhancing cross-modal feature fusion. Through the use of Contrastive Sample Distillation (CSD) and Similarity-based Representation Distillation (SRD), the framework ensures robust representation learning, even under incomplete data conditions. This work represents a significant step forward in MSA research, providing a scalable and effective solution for real-world challenges.