Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2026, 15(6), 1391-1407
Published Online November 2026 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2026.156126

基于多尺度融合与一致性优化的鲁棒多模态情感分析研究

王严仁，徐长波

北京印刷学院信息工程学院，北京

收稿日期：2026年9月1日；录用日期：2026年10月28日；发布日期：2026年11月5日

摘　要

多模态情感分析在处理文本、音频与视觉异构模态时会产生局部特征易流失、非语言线索利用不足以及跨模态融合噪声干扰等问题，这会显著削弱模型在复杂情感场景下的判别能力与鲁棒性。针对这些问题，本文提出一种基于多尺度融合与一致性优化的鲁棒多模态情感分析模型。在特征提取阶段，模型结合DeBERTa的预训练语言模型提取文本全局语义，利用两层级联1D-CNN捕获音频与视觉局部动态特征，完成了多尺度特征提取。在跨模态融合阶段，通过联合序列拼接与动态自适应掩码，经过多层Transformer在文本语境引导下实现高信噪比的跨模态交互。在预测阶段，构建分类－回归双分支结构，通过一致性优化约束极性与强度预测的一致性，从而提升情感预测的准确性与稳定性。实验结果显示，在CMU-MOSI和CMU-MOSEI数据集上，本文方法在Acc-2、Acc-7、F1、MAE及Corr等指标均优于多种基线模型，证明了该框架具有良好的识别性能与鲁棒性。

关键词

多模态情感分析，情感预测，多尺度特征提取，跨模态融合，一致性优化

Research on Robust Multimodal Sentiment Analysis Based on Multi-Scale Fusion and Consistency Optimization

Yanren Wang, Changbo Xu

School of Information Engineering, Beijing Institute of Graphic Communication, Beijing

Received: September 1, 2026; accepted: October 28, 2026; published: November 5, 2026

Abstract

Multimodal sentiment analysis encounters issues such as the loss of local features, insufficient

文章引用: 王严仁, 徐长波. 基于多尺度融合与一致性优化的鲁棒多模态情感分析研究[J]. 人工智能与机器人研究, 2026, 15(6): 1391-1407. DOI: 10.12677/airr.2026.156126

utilization of non-linguistic cues, and noise interference in cross-modal fusion when dealing with heterogeneous modalities like text, audio, and vision. These problems significantly weaken the model's discriminative ability and robustness in complex sentiment scenarios. To address these issues, this paper proposes a robust multimodal sentiment analysis model based on multi-scale fusion and consistency optimization. In the feature extraction stage, the model combines the pre-trained language model DeBERTa to extract global semantic information from text and utilizes a two-layer cascaded 1D-CNN to capture local dynamic features from audio and vision, achieving multi-scale feature extraction. In the cross-modal fusion stage, through joint sequence concatenation and dynamic adaptive masking, high signal-to-noise ratio cross-modal interaction is achieved under the guidance of text context via multiple layers of Transformer. In the prediction stage, a classification-regression dual-branch structure is constructed, and consistency optimization is used to constrain the consistency of polarity and intensity predictions, thereby enhancing the accuracy and stability of sentiment prediction. Experimental results show that on the CMU-MOSI and CMU-MOSEI datasets, our method outperforms multiple baseline models in terms of Acc-2, Acc-7, F1, MAE, and Corr metrics, demonstrating the framework's excellent recognition performance and robustness.

Keywords

Multimodal Sentiment Analysis, Sentiment Prediction, Multi-Scale Feature Extraction, Cross-Modal Fusion, Consistency Optimization

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

随着互联网、短视频平台和社交媒体的发展，用户生成各种视频数据量大幅增长。为了提升用户的体验，多模态情感分析的相关研究备受关注，已经拓宽到各种领域。相关研究已从早期的粗粒度情感极性分类，逐步拓展到情感强度建模、缺失模态鲁棒识别、动态模态选择以及语义增强情感理解等方向[1]。然而，该领域仍面临两类挑战：一是如何从不同模态中提取具有判别力的情感表征，二是如何实现高效且稳定的跨模态融合。

现有的研究方法在特征增强方面，通过语言中心建模、音视频语义补充和高效长序列建模等方式，提高单模态情感线索的表达能力[2]；在跨模态融合方面，研究者尝试利用全局到局部交互、动态路径选择和文本中心双流结构，增强不同模态之间的信息交换[3]；在表征解耦方面，相关工作通过特征重构、单模态监督、互信息约束和因子化建模，降低冗余信息与模态冲突对融合结果的影响[4] [5]；在鲁棒学习方面，缺失模态重建、不完整数据建模、模态翻译、信息瓶颈和不确定性感知路由等方法进一步提升了模型在复杂场景下的稳定性[6]。然而，现有方法多偏重全局融合或单阶段交互，容易忽略文本全局语义与音视频局部瞬时线索在表达尺度上的差异，导致情感信息容易被忽视。同时，分类极性与连续情感强度之间缺少一致性约束，模型在弱情感和模糊边界样本上容易出现预测偏移。

针对上述问题，本文提出基于多尺度融合与一致性优化的鲁棒多模态情感分析模型 CMI。特征提取阶段，采用 DeBERTa 挖掘文本全局深层语义，利用两层级联一维卷积网络建模音、视觉模态语调、表情等局部瞬时时序特征，实现多尺度特征表达。模型将三类模态特征映射至统一维度，沿时间维度拼接得到跨模态序列；构建动态自适应掩码屏蔽填充噪声，使 Transformer 仅对有效信息交互，并以文本语义为

引导对音视频特征重加权，挖掘模态间互补情感信息。最后基于融合全局表征搭建分类–回归双分支，分别预测情感极性与连续情感强度，借助多任务一致性损失约束输出，提升情感预测的准确性与鲁棒性。

本文的主要研究贡献如下：

- 我们提出一种基于多尺度融合与一致性优化的多模态情感分析模型 CMI。该模型结合 DeBERTa 与两层级联 1D-CNN，分别提取文本全局深层语义和音视频局部瞬时动态特征，实现了不同模态情感信息的多尺度建模，减少了情感特征易流失的问题。
- 我们设计跨模态动态掩码融合与多任务一致性优化机制。通过跨模态序列联合拼接、动态自适应掩码和多层 Transformer，实现文本语境引导下的高信噪比跨模态调制融合；同时构建分类–回归双分支预测结构，利用一致性损失约束情感极性与情感强度预测结果的一致性，从而提升模型在情感分析任务中的准确性与鲁棒性。
- 针对在 CMU-MOSI 和 CMU-MOSEI 两个不同的公开数据集上进行的大量的实验中表明，所提出的 CMI 模型在对情感预测中实现了最优的性能。此外，关键的模块在模型中表现出了良好的泛化能力。

2. 相关工作

2.1. 多模态特征融合与多尺度建模

在多模态情感分析中，如何高效融合异构特征并捕获跨模态的长距离依赖，始终是研究的核心。近年来，基于 Transformer 的跨模态注意力逐渐成为主流，尤其在确立文本主导地位方面取得了明显进展。TETFN [7]通过将音视频特征向文本语义空间投影，有效提升了融合表征的情感丰富度；ALMT [8]则借助语言特征引导超模态学习，抑制非语言输入中的无关噪声。针对 Transformer 二次复杂度的问题，部分研究开始引入线性复杂度的状态空间模型，MSAmba [9]和 TF-Mamba [10]分别在高效交互与缺失模态长序列融合上进行了有益探索。不过，现有框架大多默认各模态在同一时间尺度上对齐，忽略了情感表达固有的多尺度特性——文本往往是高层全局的，而面部微表情、语音突变则是局部瞬时的。这种全局融合容易平滑掉情感判断所需的关键动态线索。如何在特征提取与融合早期引入多尺度机制，成为有待突破的方向。

2.2. 情感表征与一致性优化

随着情感计算向方向深入，多模态情感分析正从离散分类走向连续强度预测。为弥合模态鸿沟并捕捉细微情感线索，近年研究在表征解耦与对比学习方面展开了有益探索。DMD 框架[11]将多模态特征解缠为模态特有与共享表征，以剔除异构噪声；LNLN [12]则引入基于主导模态校正的抗噪机制，提升不完整数据下的稳定性；Li [13]等人提出的统一自蒸馏框架，进一步在模态缺失时维持了表征一致性。因此，研究通过提出了一种通过多尺度调制融合的 MSA 模型用来进行情感预测。

3. 方法

图 1 为本文所提 CMI 模型的整体框架，包含多尺度单模态特征提取、跨模态统一序列构建、动态掩码调制融合、多任务一致性预测四大核心模块。多尺度单模态特征提取模块借助 DeBERTa 获取文本全局深层语义，采用两层级联 1D-CNN 捕捉音、视觉模态局部瞬时动态信息，完成多尺度特征建模。跨模态统一序列构建模块将三类模态特征映射至同一维度，沿时间维度拼接生成跨模态融合序列。动态掩码调制融合模块生成自适应掩码规避填充噪声干扰，通过 Transformer 以文本为引导实现音视频特征动态加权与深度交互。多任务一致性预测模块利用融合全局表征搭建分类–回归双分支，分别预测情感极性与连

续情感强度，并引入一致性损失约束两者输出，提升情感预测的准确性与鲁棒性。

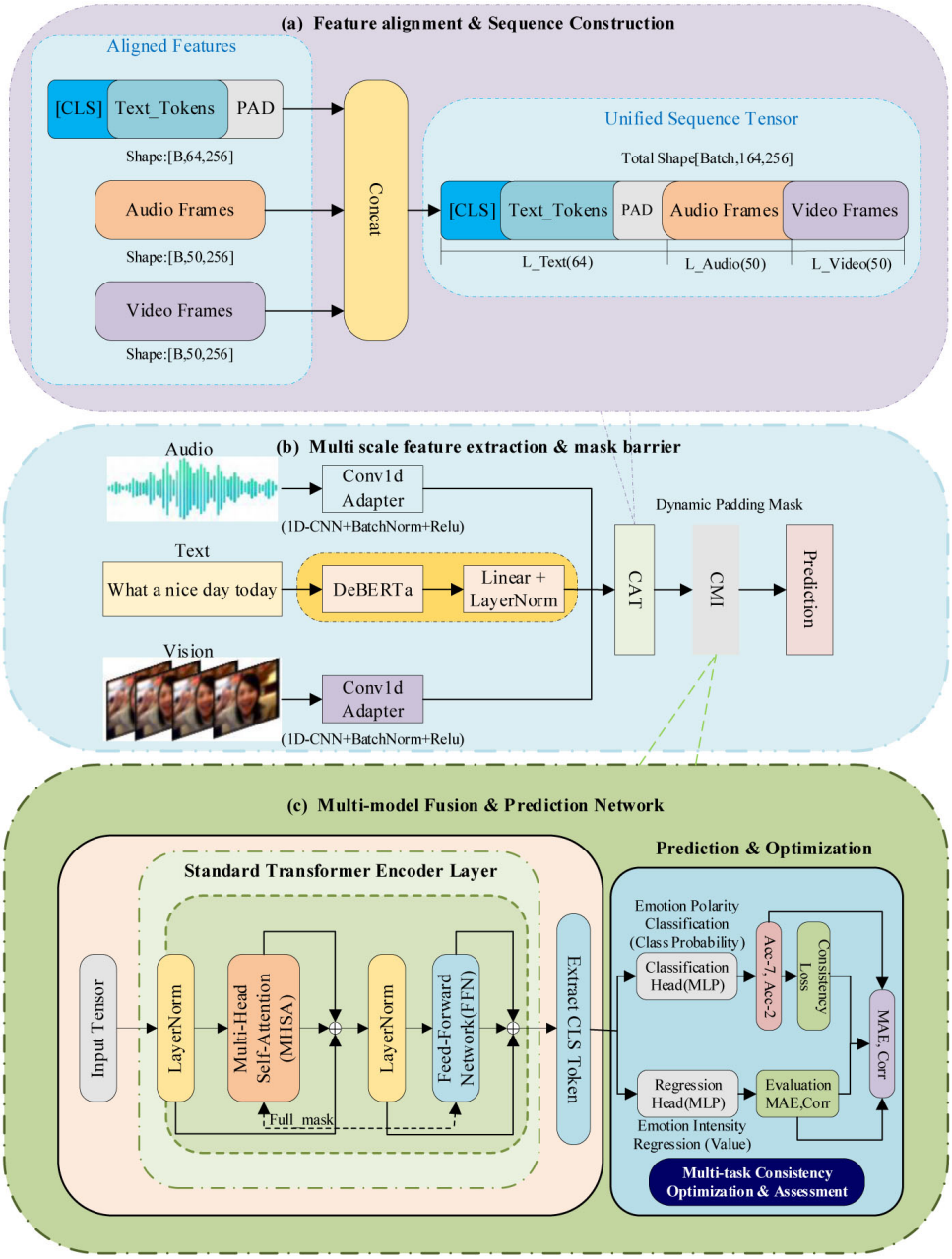


Figure 1. Illustrates the overall architecture of the research methodology in this paper. (a) Represents the framework for feature alignment and unified sequence construction; (b) Represents the framework for multi-scale feature extraction and mask barrier; (c) Represents the framework for cross-modal fusion and prediction

图 1. 本文研究方法的整体架构。(a) 是特征对齐与统一序列构建框架；(b) 是多尺度特征提取与掩码屏障框架；(c) 是跨模态融合与预测框架

3.1. 多尺度单模态特征提取

针对多模态数据存在的高维异构特性，特征提取层必须兼顾文本的高层全局语义与声画信号的局部瞬时动态。

3.1.1. 文本模态深层语义特征提取

对于文本模态，本文采用 DeBERTa-v3-base 作为文本编码器，以提取具有上下文感知能力的深层语义表征。给定文本 token 输入矩阵 $\boldsymbol{X}_t$ 及其注意力掩码 M_t，DeBERTa 的输出表示为：

$$H_{bert} = \text{DeBERTa}\left(\boldsymbol{X}_t, M_t\right) \tag{1}$$

式中，$H_{bert} \in \mathbb{R}^{B\times L_t \times d_{bert}}$，表示了模型最后一层输出的隐藏状态张量，$d_{bert}$ 表示 DeBERTa 输出的隐藏层维度为 768。

为了消除文本特征与音视频特征之间的维度差异，本文引入线性投影层将文本特征映射到统一隐藏维度 d：

$$\tilde{H}_t = H_{bert}W_t + b_t \tag{2}$$

其中，$W_t \in \mathbb{R}^{d\times d_{bert}}$，$b_t \in \mathbb{R}^d$。随后，采用层归一化增强训练稳定性：

$$H_t = \text{LayerNorm}\left(\tilde{H}_t\right) \tag{3}$$

最终得到全局语义表示，$H_t \in \mathbb{R}^{B\times L_t \times d}$。由于 DeBERTa 能够在完整句子上下文中建模词元之间的长距离依赖，因此 H_t 表示文本模态的全局语义尺度信息。

3.1.2. 音视频模态局部动态特征提取

不同于文本模态的离散语义结构，音频与视觉信号通常表现为连续时序序列，其中包含语调突变、停顿、面部表情变化和动作幅度变化等局部动态线索。为捕获此类短时局部模式，本文分别对音频和视觉模态构建结构相同的两层级联一维卷积网络。

对于任意非语言模态 $m \in \{a, v\}$，其输入特征为：$X_m \in \mathbb{R}^{B\times L_m \times d_m}$。其中，$m = a$ 表示音频模态，$m = v$ 表示视觉模态，L_m 为对应模态的序列长度，d_m 为原始特征维度。一维卷积要求通道维位于第二维，因此首先对输入进行转置：

$$\tilde{X}_m = \text{Transpose}\left(X_m\right) \in \mathbb{R}^{B\times L_m \times d_m} \tag{4}$$

第一层卷积用于捕获短时局部动态模式：

$$Z_m^{(1)} = \text{Dropout}\left(\Phi\left(\text{BN}_1\left(\text{Conv1D}_{k_1}\left(\bar{X}_m W_m^{(1)}, b_m^{(1)}\right)\right)\right)\right) \tag{5}$$

其中，Conv1D_{k_1} 表示卷积核大小为 k_1 的一维卷积操作，$W_m^{(1)}$ 和 $b_m^{(1)}$ 分别表示第一层卷积的权重和偏置，$\text{BN}_1(\cdot)$ 表示一维批归一化，$\phi(\cdot)$ 表示非线性激活函数，本文采用 ReLU。

第二层卷积进一步扩大局部感受野，用于建模更高阶的局部时序变化：

$$Z_m^{(2)} = \text{Dropout}\left(\Phi\left(\text{BN}_2\left(\text{Conv1D}_{k_2}\left(\bar{Z}_m W_m^{(2)}, b_m^{(2)}\right)\right)\right)\right) \tag{6}$$

由于卷积输出通道数已经设置为统一隐藏维度 d，经过转置后得到：

$$H_m = \text{Transpose}\left(Z_m^{(2)}\right) \in \mathbb{R}^{B\times L_m \times d} \tag{7}$$

经过上述过程，最终输出音频和视觉模态分别是：$H_a \in \mathbb{R}^{B\times L_a \times d}$，$H_v \in \mathbb{R}^{B\times L_v \times d}$。

3.2. 文本引导的多尺度调制机制

为了避免简单拼接导致不同模态信息在融合前缺少语义约束，本文进一步设计文本引导的多尺度调制机制，使非语言模态中的局部线索能够在文本语境约束下参与融合。

首先，从文本表示 H_t 中提取全局语义锚点。由于文本序列首位对应[CLS]位置，因此定义为：

$$c_t = H_t[:,0,:] \in \mathbb{R}^{B\times d} \tag{8}$$

为了使该全局语义向量能够与音频和视觉序列逐时间步交互，将其沿时间维度扩展：

$$C_a = \mathrm{Repeat}(c_t, L_a) \in \mathbb{R}^{B\times L_a\times d} \tag{9}$$

$$C_v = \mathrm{Repeat}(c_t, L_v) \in \mathbb{R}^{B\times L_v\times d} \tag{10}$$

对于音频模态，本文根据文本全局语义与音频局部特征共同生成调制门控权重：

$$G_a = \sigma([H_a, C_a])W_a^g + b_a^g \tag{11}$$

其中，$[H_a, C_a]$表示在特征维度上的拼接操作，$[H_a, C_a] \in \mathbb{R}^{B\times L_a\times 2d}$，$W_a^g \in \mathbb{R}^{d\times 2d}$，$b_a^g \in \mathbb{R}^d$表示可学习参数，$\sigma(\cdot)$表示 Sigmoid 激活函数，$G_a \in \mathbb{R}^{B\times L_a\times d}$表示音频模态的文本引导门控矩阵。随后，利用门控权重对音频特征进行调制：

$$\hat{H}_a = H_a + G_a \odot H_a \tag{12}$$

其中，$\odot$表示逐元素乘法。该残差式调制方式既保留了原始音频特征，又能够突出与当前文本语义相关的声学线索。同理，对于视觉模态，文本引导的调制过程定义为：

$$G_v = \sigma([H_v, C_v])W_v^g + b_v^g \tag{13}$$

$$\hat{H}_v = H_v + G_v \odot H_v \tag{14}$$

其中，$W_v^g \in \mathbb{R}^{d\times 2d}$，$b_v^g \in \mathbb{R}^d$。经过该过程，文本全局语义尺度与音视频局部动态尺度之间建立了显式调制关系。

本文中的多尺度建模体现在两个方面。首先，文本模态通过 DeBERTa 建模句子级上下文依赖，主要提供全局语义尺度信息；音频和视觉模态通过级联一维卷积提取短时局部动态信息，主要提供局部时序尺度信息。其次，两层卷积结构逐步扩大音视频模态的有效感受野。当卷积步长为 1 且无空洞卷积时，其有效感受野可表示为：

$$R = 1 + (k_1 - 1) + (k_2 - 1) \tag{15}$$

其中，k_1和k_2分别为两层一维卷积的卷积核大小。例如，当$k_1 = k_2 = 3$时，局部感受野为$R = 5$，模型能够同时捕获相邻帧及其上下文中的局部动态变化。由此，文本的全局语义尺度与音视频的局部动态尺度形成互补，为后续跨模态调制融合提供基础。

动态跨模态有效位置掩码构建

由于直文本、音频和视觉模态具有不同的序列长度，在批量训练过程中通常需要对各模态序列进行 padding。若不对 padding 位置进行屏蔽，Transformer 的注意力机制可能会将无效位置纳入注意力计算，从而引入噪声。为此，本文构建跨模态动态有效位置掩码，用于同时屏蔽文本、音频和视觉模态中的无效位置：

对于文本模态，其有效位置掩码由分词器直接生成：

$$M_t(i,j) = \begin{cases} 1, & \text{if the } j\text{-th token of the } i\text{-th sample is valid} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

对于音频和视觉模态，预处理阶段保留了真实序列长度l_i^a和l_i^v，其有效掩码定义为：

$$M_a(i,j) = \begin{cases} 1, & j < l_i^a \\ 0, & j \geq l_i^a \end{cases} \tag{17}$$

$$M_v(i,j)=\begin{cases}1, & j<l_i^v\\ 0, & j\geq l_i^v\end{cases} \tag{18}$$

在获得三种模态的有效位置掩码后，将其在时间维度上拼接：

$$M=\left[M_t;M_a;M_v\right]\in(0,1)^{B\times L} \tag{19}$$

其中，$L=L_t+L_a+L_v$。为适配 Transformer 中的键填充掩码，将有效位置掩码转换为无效位置掩码：

$$P=1-M \tag{20}$$

其中，$P(i,j)=1$ 表示第 i 个样本中第 j 个位置为无效位置。在多头注意力计算中，给定查询矩阵 $\boldsymbol{Q}$、键矩阵 $\boldsymbol{K}$ 和值矩阵 $\boldsymbol{V}$，注意力得分为：

$$S=\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_h}} \tag{21}$$

式中，d_h 为单个注意力头的维度。为了屏蔽无效位置，构造加性注意力掩码：

$$M_{i,j}=\begin{cases}0, & M(i,j)=1\\ -\infty, & M(i,j)=0\end{cases} \tag{22}$$

掩码后的注意力权重计算为：

$$\alpha=\mathrm{Softmax}(S+A) \tag{23}$$

最终注意力输出为：

$$O=\alpha V \tag{24}$$

为进一步避免无效 query 位置在后续层中传播，本文对 Transformer 输出再次施加有效位置约束：

$$O=O\odot M[:,:,\mathrm{None}] \tag{25}$$

通过上述方式，文本、音频和视觉模态中的 padding 位置均被显式屏蔽，避免了无效帧和补齐 token 对跨模态注意力分布的干扰。

3.3. 动态掩码约束下的跨模态联合自注意力融合

在获得文本表示 H_t、调制后的音频表示 $\hat{H}_a$ 和调制后的视觉表示 $\hat{H}_v$ 后，本文将三种模态在时间维度上进行拼接，构建统一的跨模态序列：

$$H^0=\left[H_t;\hat{H}_a;\hat{H}_v\right]\in\mathbb{R}^{B\times L\times d} \tag{26}$$

其中，$L=L_t+L_a+L_v$。为保留序列位置信息和模态来源信息，进一步加入位置嵌入和模态嵌入：

$$\bar{H}^0=H^0+E^{pos}+E^{mod} \tag{27}$$

其中，E^{pos} 表示位置嵌入，E^{mod} 表示模态类型嵌入。对于第 l 层 Transformer，输入为 H^{l-1}。首先通过线性映射得到查询、键和值矩阵：

$$Q_h^l=H^{l-1}W_h^{Q,l} \tag{28}$$

$$K_h^l=H^{l-1}W_h^{K,l} \tag{29}$$

$$V_h^l=H^{l-1}W_h^{V,l} \tag{30}$$

其中，h 表示第 h 个注意力头，$W_h^{Q,l}$、$W_h^{K,l}$、$W_h^{V,l}$ 为可学习参数矩阵。结合动态掩码后的注意力计算为：

$$\alpha_h^l = \mathrm{Softmax}\left(\frac{Q_l^{\mathrm{T}}\left(K_h^l\right)^{\mathrm{T}}}{\sqrt{d_k}} + A\right) \tag{31}$$

$$O_h^l = \alpha_h^l V_h^l \tag{32}$$

经过多个注意力头的输出拼接后进行线性投影：

$$O^l = \mathrm{Concat}\left(O_h^l, O_h^l, \cdots, O_h^l\right) W^O \tag{33}$$

其中，h 为注意力头数量，W^O 为输出投影矩阵。随后采用残差连接和层归一化：

$$\tilde{H}^1 = \mathrm{LayerNorm}\left(L^{l-1} + O^l\right) \tag{34}$$

前馈网络进一步增强非线性表达能力：

$$F^l = W_2^l \Phi\left(W_1^l \tilde{H}^1 + b_1^l\right) + b_2^l \tag{35}$$

$$H^l = \mathrm{LayerNorm}\left(\tilde{H}^1 + F^l\right) \tag{36}$$

经过 N 层 Transformer 后，得到最终融合表示：

$$H^f = H^N \in \mathbb{R}^{B\times L\times d} \tag{37}$$

该联合自注意力机制能够在统一序列空间中建模文本、音频和视觉之间的全局依赖关系。由于融合前已经通过文本语义对音视频局部特征进行了调制，且融合过程中引入了跨模态动态掩码，因此模型能够在抑制无效 padding 干扰的同时，实现高信噪比的跨模态情感信息交互。

3.4. 情感预测与多任务一致性优化目标

为实现高精度的分析与离散情感极性判断，研究中在网络的末端构建了并行的决策系统与多任务联合一致性优化目标。

3.4.1. 向量级门控与并行预测分支

在经过多层 Transformer 跨模态融合后，模型获得高阶多模态融合表示。由于联合序列的首位特征对应文本全局语义锚点，并且已经在跨模态交互过程中融合了音频和视觉模态中的情感信息，因此本文提取该位置的全局融合表征作为最终预测的输入。对于第 i 个样本，其全局融合特征可表示为：

$$h_i = H_i^{fuse}[0] \tag{38}$$

其中，H_i^{fuse} 表示第 i 个样本经过 Transformer 融合后的多模态序列表示，$H_i^{fuse}[0]$ 表示序列首位的全局融合向量，$h_i \in \mathbb{R}^d$，d 为隐藏层维度。为了进一步突出与情感预测相关的关键信息，并抑制融合表征中的冗余或噪声特征，本文在预测前引入向量级门控机制。该机制通过自适应权重对全局融合特征进行逐维筛选：

$$g_i = \sigma\left(W_g h_i + b_g\right) \tag{39}$$

$$\bar{h}_i = g_i \odot h_i \tag{40}$$

其中，W_g 和 b_g 分别表示门控层的可学习权重和偏置，$\odot$ 表示逐元素乘法，g_i 表示向量级门控权重，$\bar{h}_i$ 表示经过门控筛选后的情感增强融合表征。之后模型构建分类－回归两个并行预测分支。回归分支用于预测连续情感强度，其输出为：

$$\hat{y}_i^{reg} = f_{reg}\left(\bar{h}_i\right) = W_{r2}\phi\left(W_{r1}\bar{h}_i + b_{r1}\right) + b_{r2} \tag{41}$$

其中，$\hat{y}_i^{reg}$ 表示第 i 个样本的连续情感强度预测值，取值空间对应数据集中的连续情感标签区间；W_{r2}、W_{r1}、b_{r1} 和 b_{r2} 为回归分支中的可学习参数，$\phi(\cdot)$ 表示非线性激活函数。

分类分支用于预测离散情感类别，其输出为分类 logits 向量：

$$Z_i^{cls} = f_{cls}\left(\bar{h}_i\right) = W_{c2}\phi\left(W_{c1}\bar{h}_i + b_{c1}\right) + b_{c2} \tag{42}$$

其中，$Z_i^{cls} = \left[Z_i^1, Z_i^2, \cdots, Z_i^K\right]$，表示第 i 个样本在 K 个情感类别上的分类逻辑值；W_{c2}、W_{c1}、b_{c1} 和 b_{c2} 为分类分支中的可学习参数。通过分类–回归双分支结构，模型能够同时完成离散情感极性判别和连续情感强度估计，为后续一致性优化提供基础。

3.4.2. 联合一致性优化目标

为使模型同时具备连续情感强度预测能力和离散情感类别判别能力，本文在分类–回归双分支预测结构的基础上，设计了多任务联合一致性优化目标。

给定一个批次中第 i 个样本的真实连续情感标签为 y_i，回归分支输出的连续情感预测值为 $\hat{y}_i^{reg}$，分类分支输出的 logits 向量为：

$$Z_i^{cls} = \left[Z_i^1, Z_i^2, \cdots, Z_i^K\right] \tag{43}$$

其中，K 表示情感类别数。在七分类情感预测任务中，$K = 7$，对应的离散情感类别中心为：

$c = [-3, -2, -1, 0, 1, 2, 3]$

首先，本文采用平均绝对误差作为主回归损失，来衡量回归分支输出的连续情感强度与真实标签之间的偏差：

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i^{reg} - y_i\right| \tag{44}$$

其中，N 表示批次样本数量，y_i 表示真实连续情感标签，$\hat{y}_i^{reg}$ 表示回归分支预测得到的连续情感强度值。

对于分类分支，需要先将连续标签 y_i 转换为离散类别标签。由于 CMU-MOSI 和 CMU-MOSEI 数据集中的情感强度标签通常位于[−3, 3]区间，本文通过四舍五入与截断操作将其映射为七分类标签：

$$q_i = \mathrm{clip}\left(\mathrm{round}\left(y_i\right), -3, 3\right) + 3 \tag{45}$$

其中，$\mathrm{round}(\cdot)$ 表示四舍五入操作，$\mathrm{clip}(\cdot)$ 表示将标签限制在[−3, 3]范围内，最后加 3 是为了将情感类别{−3, −2, −1, 0, 1, 2, 3}转换为分类索引{0, 1, 2, 3, 4, 5, 6}。

随后，对分类分支输出的 logits 进行 Softmax 归一化，得到第 i 个样本属于第 k 个情感类别的概率：

$$p_i^k = \frac{\exp\left(z_i^k\right)}{\sum_{j=1}^{K}\exp\left(z_i^j\right)} \tag{46}$$

其中，z_i^k 表示第 i 个样本在第 k 个情感类别上的分类输出，p_i^k 表示对应类别的预测概率。分类损失采用交叉熵损失表示为：

$$\mathcal{L}_{cls} = -\frac{1}{N}\sum_{i=1}^{N}\log p_i^{q_i} \tag{47}$$

为了进一步增强分类分支与回归分支之间的预测一致性，本文将分类概率分布映射回连续情感空间。根据分类概率分布和情感类别中心向量 c，计算分类分支对应的连续情感期望值：

$$\tilde{y}_i = \sum_{k=1}^{K} p_i^k c_k \tag{48}$$

其中，c_k 表示第 k 个离散情感类别对应的情感强度中心，$\tilde{y}_i$ 表示由分类概率分布计算得到的连续情感期望值。在此基础上，本文定义分类回归一致性约束损失为：

$$\mathcal{L}_{Cons}=\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}_i^{reg}-\tilde{y}_i\right)^2 \tag{49}$$

该损失项通过最小化回归分支连续预测值 $\hat{y}_i^{reg}$ 与分类概率期望值 $\tilde{y}_i$ 之间的差异，使分类预测与回归预测在情感方向和强度表达上保持一致。对于弱情感、中性情感以及正负情感边界不明显的样本，该约束能够有效减少两类预测结果之间的偏移，从而提升模型的预测稳定性。

最终，多任务联合优化目标定义为：

$$\mathcal{L}_{total}=\mathcal{L}_{reg}+\alpha\mathcal{L}_{cls}+\lambda\mathcal{L}_{Cons} \tag{50}$$

其中，$\mathcal{L}_{reg}$ 表示连续情感强度回归损失，$\mathcal{L}_{cls}$ 表示离散情感类别分类损失，$\mathcal{L}_{Cons}$ 表示分类–回归一致性约束损失；α 和 λ 分别为分类损失和一致性损失的权重系数。

4. 实验

4.1. 数据集和评估指标

4.1.1. 实验数据集

本文选取了 CMU-MOSEI 与 CMU-MOSI 两个基准数据集来评估所提模型有效性与预测鲁棒性。CMU-MOSEI 数据集包含了超过 23,453 个标注的视频片段。在特征层面，每个样本均包含了词元级的文本输入、由 COVAREP 提取的 74/80 维声学特征，以及由 FACET 提取的 35 维视觉面部特征。CMU-MOSI 数据集共包含 2199 个视频片段，涵盖 89 名不同的演讲者。两个基准数据集都标注为在[−3, 3]区间的连续实数值。

4.1.2. 评估指标

为了衡量模型在连续情感强度预测以及离散情感极性分类上的综合能力，研究中采用了与现有多模态情感分析评价体系严格对齐的多种指标。具体的评估指标定义如下：

平均绝对误差(MAE)：衡量模型预测的连续情感分值 $\hat{y}$ 与真实标注 y 之间的绝对差异，是回归任务的核心指标。计算公式为：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}\left|\hat{y}_i-y_i\right| \tag{51}$$

式中，N 表示验证集或测试集中的总体样本数量；$\hat{y}_i\in[-3,3]$ 为第 i 个样本的真实人类情感标注强度；$\hat{y}_i$ 则是模型回归头输出的预测连续值，该值越小，说明模型对情感强度的拟合误差越低。

皮尔逊相关系数(Corr)：为了评估模型对情感强度变化趋势的捕捉能力，系统计算了预测分布与真实分布之间的线性相关性。该指标通过协方差与标准差的商值求得：

$$\mathrm{Corr}=\frac{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)\left(\hat{y}_i-\bar{\hat{y}}_i\right)}{\sqrt{\sum_{i=1}^{N}\left(y_i-\bar{y}\right)^2\sum_{i=1}^{N}\left(\hat{y}_i-\bar{\hat{y}}_i\right)^2}} \tag{52}$$

式中，$\bar{y}$ 与 $\bar{\hat{y}}$ 分别代表真实标签集和预测标签集的算术平均值。Corr 的取值范围在[−1, 1]之间，数值越接近 1，意味着模型预测的情感起伏与人类真实情感波动高度同频。

七分类准确率(Acc-7)：面对高度细分的情感等级，不仅需要预测连续值，还需将其归态化为离散类别。在研究中将回归模型的连续预测值 $\hat{y}_i$ 通过四舍五入(Round)和截断操强制映射至七个离散情感等级

{−3, −2, −1, 0, 1, 2, 3}中。具体的分类判定公式为：

$$\text{Acc-7} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\max\left(-3, \min\left(3, \text{Round}\left(\hat{y}_i\right)\right)\right) = y_i\right) \tag{53}$$

其中，$\mathbb{I}(\cdot)$ 为指示函数，当内部逻辑表达式成立时取值为 1，否则为 0。

包含中性类的二分类准确率(Acc-2_{has0})：该指标将中性情感统一归入正类边界。本文通过构建布尔型预测器，若预测值大于等于 0 则记为 1 (正极性)，否则记为 0 (负极性)。计算公式为：

$$\text{Acc-2}_{has0} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(\mathbb{I}\left(\hat{y}_i \geq 0\right) = \mathbb{I}\left(y_i \geq 0\right)\right) \tag{54}$$

式中，外层的 $\mathbb{I}(\cdot)$ 为匹配指示函数，当预测的极性布尔值与真实标注的极性布尔值完全一致时取 1，否则取 0；N 为全量测试样本数。

剔除中性类的二分类准确率(Acc-2_{non0})为了排除模棱两可的中性情感对极性判断的干扰，在计算前使用非零掩码。设非零样本的索引集合为 $S_{non0} = \{i \mid y_i \neq 0\}$，且该子集的样本总数为 N_{non0}，仅对该子集内的样本进行匹配计算：

$$\text{Acc-2}_{non0} = \frac{1}{N_{non0}}\sum_{i \in S_{non0}}\mathbb{I}\left(\mathbb{I}\left(\hat{y}_i > 0\right) = \mathbb{I}\left(y_i > 0\right)\right) \tag{55}$$

该式精确剔除了 $y_i = 0$ 时的模糊地带。同时，考虑到数据集中正负类别的潜在不平衡，本文采用了基于样本真实占比的加权 F_1：

$$F_1 = \sum_{c \in (0,1)} w_c \cdot \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c} \tag{56}$$

其中，w_c 表示类别 c 在非零子集中的样本占比。

4.2. 基线模型

为了验证所提出模型的有效性，本文与多个基线以及最新的方法进行了公平的比较。

TFN [14]：张量融合网络 TFN 通过张量外显式捕获所有单模态、双模态和三模态特征完全交互信息。

LMF [15]：通过引入低秩张量分解技术，在不构高维特征张量的情况下，可以提取多模态间的交互信息。

MIMM [16]：MIMM 针对多模态融合易出现信息丢失问题，以互信息最大化为正则化约束，在特征融合全过程中保留单模态关键情感线索，提升情感分析效果。

Self-MM [17]：Self-MM 采用自监督生成单模态标签来辅助各模态学习自身情感表示，解决了多模态情感数据集只有全局融合标签、无单模态标签的问题。

Tri-MCA [18]：Tri-MCA 设计联合三模态交叉注意力，搭配维度门控与动态模态权重，同时建模三元交互、过滤噪声、平衡模态贡献，解决传统两两注意力、静态融合无法适配不均衡多模态信号的缺陷。

PFGAD [19]：PFGAD 以文本渐进融合获取联合特征，引入双辅助分支约束 VIB 降噪，平衡信息压缩与有效情感特征留存，解决单一信息瓶颈易过度降噪、浅层融合交互不足的问题。

CMSA [20]：CMSA 用文本校正消除视听语义冲突，搭配上下文 MoE 适配不同场景，动态调整模态表征，解决模态信息矛盾、同文本随场景情感歧义两大多模态分析难题。

4.3. 情感预测的鲁棒性分析

为验证所提模型的有效性，本文在固定数据划分下，采用 5 个不同的随机种子对所提模型进行独立

训练，并在测试集上进行评估。表 1 是本文模型与现有多模态情感分析基准方法在 CMU-MOSI 和 CMU-MOSEI 数据集上的性能对比，模型的实验结果 5 次独立实验的均值 ± 标准差表示。根据表中的结果，本文模型在两个公开数据集上的平均性能总体优于其他对比方法，最优结果以粗体表示。

Table 1. The performance comparison between the proposed method and the baseline method on the datasets
表 1. 本文方法与基线方法在数据集上的性能比较

Models	CMU-MOSI					CMU-MOSEI				
	F1↑	Acc-2↑	Acc-7↑	MAE↓	Corr↑	F1↑	Acc-2↑	Acc-7↑	MAE↓	Corr↑
TFN	80.7	80.8	34.9	0.901	0.698	82.1	82.5	50.2	0.593	0.700
LMF	82.4	82.5	33.2	0.917	0.695	82.1	82.0	48.0	0.623	0.677
MIMM	85.9	86.0	46.6	0.712	0.800	85.9	85.9	54.2	0.526	0.722
Self-MM	84.9	84.8	44.9	0.713	0.798	85.3	85.1	53.8	0.530	0.765
Tri-MCA	84.5	85.6	48.1	0.701	0.802	86.5	86.9	56.1	0.518	0.781
PFGAD	87.5	87.5	46.2	0.694	0.814	86.7	86.7	53.8	0.554	0.786
CMSA	87.1	87.3	-	0.698	0.822	85.4	85.7	-	0.502	0.788
Ours	**93.03 ± 0.13**	**92.89 ± 0.24**	**48.60 ± 0.11**	**0.595 ± 0.015**	**0.871 ± 0.006**	**88.77 ± 0.25**	**88.97 ± 0.16**	**57.59 ± 0.36**	**0.499 ± 0.006**	**0.791 ± 0.011**

由表 1 可知，本文方法在 CMU-MOSI 和 CMU-MOSEI 两个数据集上均取得了优于基线方法的整体结果。CMU-MOSI 上，CMI 的 F1、Acc-7 和 Corr 分别为 93.03% ± 0.13%、48.60% ± 0.11%和 0.871 ± 0.006，MAE 降至 0.595 ± 0.015；CMU-MOSEI 上，CMI 的 F1、Acc-7、MAE 和 Corr 分别为 88.77% ± 0.25%、57.59% ± 0.36%、0.499 ± 0.006 和 0.791 ± 0.011。CMI 结果采用 5 次独立实验的均值 ± 标准差，基线结果来自原论文，因此不对基线结果人为补充标准差。

在 CMU-MOSI 数据集上，本文方法的 F1、Acc-2、Acc-7、MAE 和 Corr 分别为 93.03% ± 0.13%、92.89% ± 0.24%、48.60% ± 0.11%、0.595 ± 0.015 和 0.871 ± 0.006。与表中最优基线 PFGAD、MIMM 或 CMSA 相比，本文方法在分类准确率、七分类准确率、连续情感强度误差和相关性等指标上均取得更优结果，说明多尺度特征提取、文本引导调制和动态掩码融合能够有效提升情感识别性能。

在 CMU-MOSEI 数据集上，本文方法的 F1、Acc-2、Acc-7、MAE 和 Corr 分别为 88.77% ± 0.25%、88.97% ± 0.16%、57.59% ± 0.36%、0.499 ± 0.006 和 0.791 ± 0.011。与表现较好的 Tri-MCA、PFGAD 和 CMSA 等基线相比，本文方法在五项指标上均保持竞争优势，尤其在 Acc-7 和 MAE 上表现突出。研究结果表明，多尺度局部动态建模与文本引导的深度融合能够同时改善离散极性判断和连续情感强度拟合。

研究结果显示，模型通过一致性损失优化与文本引导的深度调制，不仅实现了极性分类准确度提升，还在情感强度的连续拟合上展现出了极强的鲁棒性与先进性。

4.4. 消融研究

为了深入评估本文提出的多尺度调制融合模型中各个关键模块与策略的有效性，在 CMU-MOSI 和 CMU-MOSEI 数据集上分别开展了 7 组消融实验。依次移除一致性损失、Transformer 融合、局部 CNN、端到端 DeBERTa 微调、文本引导调制、动态掩码和向量级门控中的一个组件，其余设置保持不变，并比较 F1、Acc-7、MAE 和 Corr 四项核心指标。所有消融结果均为 5 次独立实验的均值 ± 标准差。

CMI w/o Consistency Loss：在该变体中，移除多任务一致性优化约束，这会引发预测值的漂移，无

法保证模型在面临模糊情感边界时的鲁棒性。

CMI w/o Transformer Fusion：在该变体中，移除跨模态多尺度调制融合机，这样丢失文本引导的动态重加权能力，而且无法利用掩码过滤补齐噪声，因此难以有效弥合高维异质数据的模态鸿沟。

CMI w/o Local CNN：在该变体中，移除声学局部特征提取模块，而音频特征只通过线性全连接层进行孤立降维，这会忽略了语调突变等强时序相关的局部动态线索，而且削弱了非语言信号的表达张力，无法为分析提供关键的情感阶跃信息。

CMI w/o E2E DeBERTa/Offline Text：通过移除模型中的预训练语言模型的在线更新机制，这样会切断了跨模态融合层对底层语义的梯度反馈，限制了特征的上下文适应性，无法精准捕捉复杂隐喻场景下的深层情感极性。

CMI w/o Text-guided Modulation：在该变体中，移除文本引导的音视频特征调制模块，但保留 DeBERTa 文本编码器、局部 CNN、动态掩码和 Transformer 跨模态融合结构。

CMI w/o Dynamic Mask：在该变体中，移除跨模态动态有效位置掩码，但保留文本引导调制、局部 CNN 和 Transformer 融合模块。

CMI w/o Vector Gate：移除向量级门控机制，使融合表示无法根据不同样本的情感特征动态调整各任务相关信息的权重。

CMI：本文提出的模型通过多尺度调制融合与一致性优化来提升多模态情感识别的准确性与鲁棒性。

Table 2. Ablation study based on the CMU-MOSEI dataset (Significant values are displayed in bold)
表 2. 基于 CMU-MOSEI 数据集的消融研究(显著值以粗体显)

	CMU-MOSEI			
	F1↑	Acc-7↑	MAE↓	Corr↑
CMI w/o Consistency Loss	83.38 ± 0.32	54.11 ± 0.19	0.637 ± 0.016	0.613 ± 0.025
CMI w/o Transformer Fusion	83.97 ± 0.11	55.95 ± 0.13	0.682 ± 0.021	0.536 ± 0.019
CMI w/o Local CNN	84.06 ± 0.14	53.12 ± 0.22	0.618 ± 0.020	0.624 ± 0.014
CMI w/o E2E DeBERTa/Offline Text	81.14 ± 0.20	55.51 ± 0.09	0.738 ± 0.011	0.657 ± 0.017
CMI w/o Text-guided Modulation	82.56 ± 0.08	50.20 ± 0.15	0.856 ± 0.014	0.596 ± 0.020
CMI w/o Dynamic Mask	81.78 ± 0.16	52.56 ± 0.24	0.763 ± 0.023	0.611 ± 0.015
CMI w/o Vector gate	83.34 ± 0.22	54.21 ± 0.16	0.803 ± 0.013	0.563 ± 0.006
CMI	**88.77 ± 0.25**	**57.59 ± 0.36**	**0.499 ± 0.006**	**0.791 ± 0.011**

由表 2 可知，在 CMU-MOSEI 数据集上，完整 CMI 模型在 F1、Acc-7、MAE 和 Corr 四项指标上均优于各消融变体。移除文本引导调制后性能下降最明显之一，说明文本全局语义对音视频局部特征的动态筛选十分重要；移除动态掩码和 Transformer 融合后，模型的跨模态交互与噪声抑制能力减弱；移除一致性损失或向量级门控后，连续回归与离散分类之间的协调性受到影响。结果验证了各组件对模型性能的独立贡献。

在此基础上，本研究在 CMU-MOSEI 数据集上对完整模型及 7 个消融变体的 Acc-7 特征进行 t-SNE 可视化，如图 2 所示。为保证图示具有可比性，各子图采用相同的数据划分、特征提取位置和可视化流程。

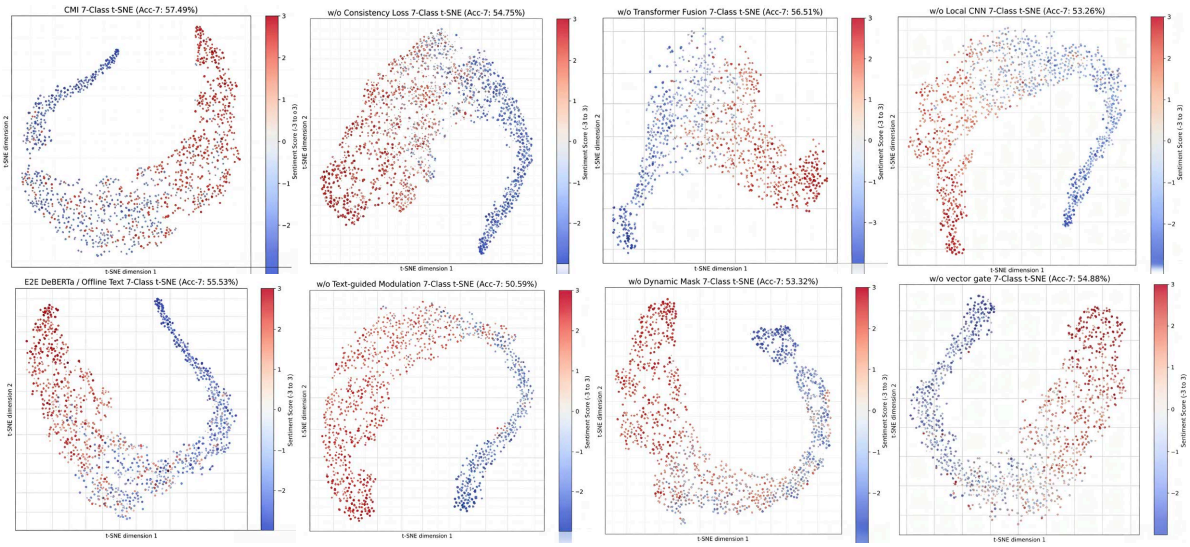


Figure 2. T-SNE visualization of the results of the ablation study on Acc-7 using the proposed model on the CMU-MOSEI dataset
图 2. 所提模型在 CMU-MOSEI 数据集上对 Acc-7 消融研究的结果进行 t-SNE 可视化

由图 2 可观察到，完整 CMI 模型的 Acc-7 为 57.59% ± 0.36%，其情感特征在低维空间中具有更清晰的类别结构，正负情感区域的过渡更加平滑。移除文本引导调制、动态掩码或 Transformer 融合后，类间边界出现不同程度的混叠，说明文本语境约束、无效位置屏蔽和跨模态交互共同提升了融合特征的可分性。

Table 3. The ablation study based on the CMU-MOSI dataset (Significant values are displayed in bold)
表 3. 基于 CMU-MOSI 数据集的消融研究(显著值以粗体显示)

	CMU-MOSI			
	F1↑	Acc-7↑	MAE↓	Corr↑
CMI w/o Consistency Loss	83.13 ± 0.15	41.84 ± 0.19	0.792 ± 0.018	0.659 ± 0.013
CMI w/o Transformer Fusion	84.24 ± 0.18	38.37 ± 0.32	0.836 ± 0.019	0.592 ± 0.024
CMI w/o Local CNN	84.98 ± 0.21	43.78 ± 0.26	0.871 ± 0.020	0.664 ± 0.017
CMI w/o E2E DeBERTa/Offline Text	82.54 ± 0.13	41.25 ± 0.12	0.936 ± 0.014	0.617 ± 0.010
CMI w/o Text-guided Modulation	83.56 ± 0.20	42.13 ± 0.17	0.875 ± 0.023	0.538 ± 0.019
CMI w/o Dynamic Mask	84.67 ± 0.16	40.76 ± 0.23	0.763 ± 0.011	0.597 ± 0.014
CMI w/o Vector gate	82.98 ± 0.10	44.63 ± 0.21	0.827 ± 0.017	0.607 ± 0.023
CMI	**93.03 ± 0.13**	**48.60 ± 0.11**	**0.595 ± 0.015**	**0.871 ± 0.006**

由表 3 可知，在 CMU-MOSI 数据集上，完整 CMI 模型的 F1、Acc-7、MAE 和 Corr 分别为 93.03% ± 0.13%、48.60% ± 0.11%、0.595 ± 0.015 和 0.871 ± 0.006，均优于相应消融变体。移除 Transformer 融合、端到端 DeBERTa 微调或一致性损失后，分类和回归指标均出现明显退化；移除文本引导调制后 Corr 降至 0.538 ± 0.019，说明文本语义引导对情感变化趋势建模尤其重要。两个数据集上的一致性结果进一步证明了 CMI 各模块的协同有效性。

为评估本文提出的多尺度调制融合模型中各核心组件的独立贡献，本文在 CMU-MOSI 和 CMU-MOSEI 数据集上开展了系统的消融研究。表 2 和表 3 分别给出了两数据集上的实验结果。

为评估本文提出的多尺度融合与一致性优化的鲁棒多模态情感分析模型 CMI 中各核心组件的独立贡献，本文通过在 CMU-MOSI 和 CMU-MOSEI 数据集上开展了消融研究。研究结果如表 3 基于 CMU-MOSI 和 CMU-MOSEI 数据集的消融研究，从表 3 的各项数据可以得出完整的 CMI 模型在 F1、Acc-7、MAE 和 Corr 四项核心指标上均获最优，任意模块缺失均会导致预测性能都有所降低。

在 CMU-MOSI 数据集上，对七分类准确率做了混淆矩阵分析，如图 3 所示的混淆矩阵。

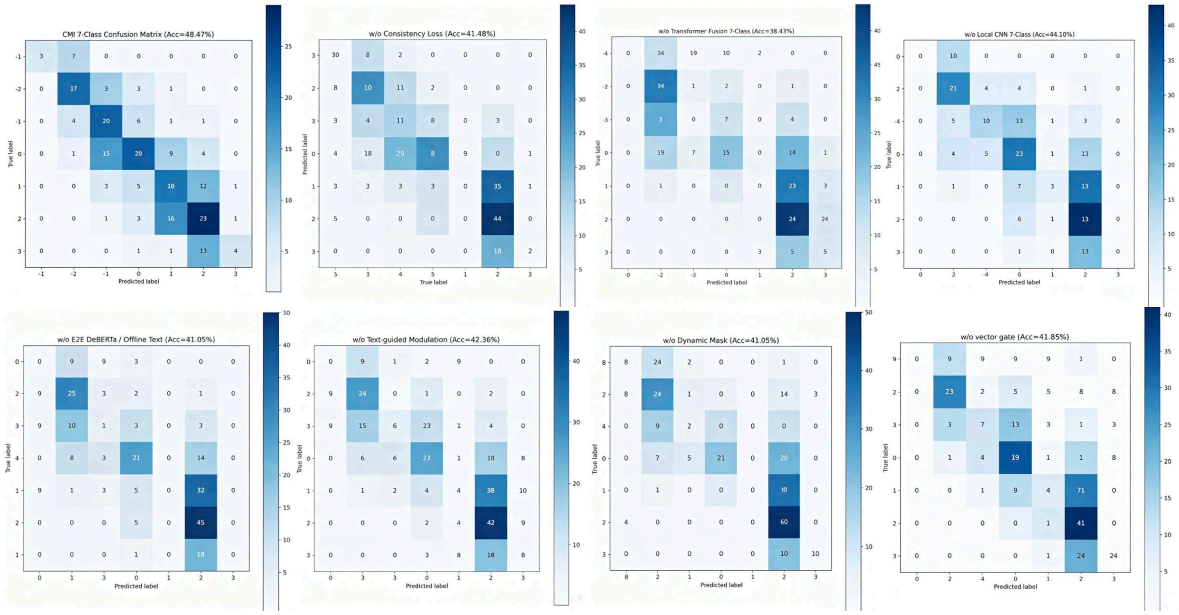


Figure 3. Confusion matrix analysis of the model's results on the Acc-7 ablation study on the CMU-MOSI dataset
图 3. 模型在 CMU-MOSI 数据集上对 Acc-7 消融研究的结果进行混淆矩阵分析

由图 2、图 3 及表 2、表 3 的消融实验结果可知，本文所提完整 CMI 模型在 CMU-MOSI、CMU-MOSEI 两个数据集上的 F1 值、Acc-7、MAE、Corr 等核心指标均取得最优性能，移除任意模块均会导致模型性能出现不同程度退化，证明各核心模块具备显著互补性。从 CMU-MOSI 数据集的混淆矩阵结果来看，完整模型的 Acc-7 可达 48.60% ± 0.11%，性能优于所有消融变体。模型对强负向、强正向情感类别的识别效果稳定，误差主要来源于相邻情感等级的混淆；去除关键模块后，相邻类别混淆问题显著加剧，印证了多尺度特征提取、跨模态融合与一致性优化可有效优化情感类别边界。

由实验表明，Transformer 跨模态融合、一致性约束优化、局部 CNN 声学特征提取、端到端动态文本微调四大核心组件，以及文本引导调制、动态掩码、向量级门控机制均对模型性能有正向增益。同时各模块的作用具有数据集针对性，文本引导调制与端到端文本表征侧重提升 CMU-MOSEI 数据集性能，Transformer 融合与文本微调则对 CMU-MOSI 分类效果至关重要。CMI 模型的性能提升并非单一模块的作用，而是多尺度表征、噪声抑制、跨模态交互与一致性决策机制协同发力的结果，有效保障了多模态情感识别的准确性与鲁棒性。

5. 结论

本文针对多模态情感分析存在的特征流失、模态异质性突出、非语言信息利用不足等问题，提出一种基于多尺度融合与一致性优化的情感预测模型。模型通过 DeBERTa 与 1D-CNN 分别对文本全局语义和音视频局部动态特征，并结合动态自适应掩码与多层 Transformer，实现文本引导的高信噪比跨模态融

合。同时引入多任务一致性优化策略，对齐分类与回归任务输出，有效提升模型预测鲁棒性。在 CMU-MOSI、CMU-MOSEI 数据集上的实验结果表明，该模型多项核心指标均有提升，各模块协同性良好，为多模态情感分析优化提供了有效方案。

参考文献

[1] Guo, Z., Jin, T. and Zhao, Z. (2024) Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. *Proceedings of the* 62*nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Bangkok, 11-16 August 2024, 1726-1736. https://doi.org/10.18653/v1/2024.acl-long.94

[2] Wu, S., He, D., Wang, X., Wang, L. and Dang, J. (2025) Enriching Multimodal Sentiment Analysis through Textual Emotional Descriptions of Visual-Audio Content. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 1601-1609. https://doi.org/10.1609/aaai.v39i2.32152

[3] Zhuang, Y., Zhang, Y., Hu, Z., Zhang, X., Deng, J. and Ren, F. (2024) GLoMo: Global-Local Modal Fusion for Multimodal Sentiment Analysis. *Proceedings of the 32nd ACM International Conference on Multimedia. Melbourne*, Melbourne, 28 October-1 November 2024, 1800-1809. https://doi.org/10.1145/3664647.3681527

[4] Fu, Y., Huang, B., Wen, Y.-J. and Zhang, P. (2024) FDR-MSA: Enhancing Multimodal Sentiment Analysis through Feature Disentanglement and Reconstruction. *Knowledge-Based Systems*, **297**, Article 111965. https://doi.org/10.1016/j.knosys.2024.111965

[5] Liu, Y. and Wang, S. (2026) FINE: Factorized Multimodal Sentiment Analysis via Mutual INformation Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, **40**, 1955-1963. https://doi.org/10.1609/aaai.v40i3.37176

[6] Luo, M., Jiang, Y. and Mai, S. (2025) Towards Explainable Fusion and Balanced Learning in Multimodal Sentiment Analysis. *Proceedings of the* 33*rd ACM International Conference on Multimedia*, Dublin, 27-31 October 2025, 1997-2006.

[7] Wang, D., Guo, X., Tian, Y., Liu, J., He, L. and Luo, X. (2023) TETFN: A Text Enhanced Transformer Fusion Network for Multimodal Sentiment Analysis. *Pattern Recognition*, **136**, Article 109259. https://doi.org/10.1016/j.patcog.2022.109259

[8] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y. and Yu, T. (2023) Learning Language-Guided Adaptive Hyper-Modality Representation for Multimodal Sentiment Analysis. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*, Singapore, 6-10 December 2023, 756-767. https://doi.org/10.18653/v1/2023.emnlp-main.49

[9] He, X., Liang, H., Peng, B., Xie, W., Khan, M.H., Song, S., *et al*. (2025) Msamba: Exploring Multimodal Sentiment Analysis with State Space Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, **39**, 1309-1317. https://doi.org/10.1609/aaai.v39i2.32120

[10] Li, X., Cheng, X., Miao, D., Zhang, X. and Li, Z. (2025) TF-Mamba: Text-Enhanced Fusion Mamba with Missing Modalities for Robust Multimodal Sentiment Analysis. *Findings of the Association for Computational Linguistics*: *EMNLP* 2025, Suzhou, 4-9 November 2025, 11252-11267. https://doi.org/10.18653/v1/2025.findings-emnlp.602

[11] Li, Y., Wang, Y. and Cui, Z. (2023) Decoupled Multimodal Distilling for Emotion Recognition. 2023 *IEEE*/*CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 6631-6640. https://doi.org/10.1109/cvpr52729.2023.00641

[12] Zhang, H., Wang, W. and Yu, T. (2024) Towards Robust Multimodal Sentiment Analysis with Incomplete Data. *Advances in Neural Information Processing Systems* 37, Vancouver, 10-15 December 2024, 55943-55974. https://doi.org/10.52202/079017-1779

[13] Li, M., Yang, D., Lei, Y., Wang, S., Wang, S., Su, L., *et al*. (2024) A Unified Self-Distillation Framework for Multimodal Sentiment Analysis with Uncertain Missing Modalities. *Proceedings of the AAAI Conference on Artificial Intelligence*, **38**, 10074-10082. https://doi.org/10.1609/aaai.v38i9.28871

[14] Zadeh, A., Chen, M., Poria, S., Cambria, E. and Morency, L. (2017) Tensor Fusion Network for Multimodal Sentiment Analysis. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September 2017, 1103-1114. https://doi.org/10.18653/v1/d17-1115

[15] Liu, Z., Shen, Y., Lakshminarasimhan, V.B., Liang, P.P., Bagher Zadeh, A. and Morency, L. (2018) Efficient Low-Rank Multimodal Fusion with Modality-Specific Factors. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), Melbourne, 15-20 July 2018, 2247-2256. https://doi.org/10.18653/v1/p18-1209

[16] Han, W., Chen, H. and Poria, S. (2021) Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, Punta Cana, 7-11 November 2021, 9180-9192. https://doi.org/10.18653/v1/2021.emnlp-main.723

[17] Yu, W., Xu, H., Yuan, Z. and Wu, J. (2021) Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 10790-10797. https://doi.org/10.1609/aaai.v35i12.17289

[18] Alrayzah, A. (2026) Tri-MCA Fusion: Cross-Modal Attention and Dynamic Gating for Multimodal Sentiment Analysis. *Scientific Reports*, **16**, Article No. 25201. https://doi.org/10.1038/s41598-026-56224-y

[19] Zheng, Z., Ye, X. and Manoharan, S. (2026) Progressive Fusion and Guidance-Aware Denoising for Robust Multimodal Sentiment Analysis. *Scientific Reports*, **16**, Article No. 27977. https://doi.org/10.1038/s41598-026-56790-1

[20] Zhao, J., Deng, O. and Jin, Q. (2025) Multimodal Sentiment Analysis CMSA: Addressing Semantic Discrepancy and Context Dependency in Multimodal Sentiment Analysis. SSRN Preprint.