

2026 年中国研究生数学建模竞赛 E 题

复杂场景下多模态情感预测的数学建模与算法设计

一、背景

情感分析是具身智能、自然人机交互等技术领域的关键支撑技术之一。真实交互中人类情绪通过文本语义、语音韵律、面部表情与肢体动作等多通道信号协同表达，多模态信息联合建模是突破单模态识别局限、提升复杂场景下情绪理解准确性与稳定性的核心路径。然而，多模态情感预测在落地中面临三重挑战：文本、语音、视觉模态在采样频率、特征维度与时间尺度上差异显著，特征提取与时序对齐的质量直接决定模型性能上限；实际采集常因画面遮挡、环境噪声、转写失败等导致模态局部缺失，常规固定权重融合模型性能急剧下降；多数模型仅能输出预测结果，决策依据难以追溯与复核，制约了模型的可信应用。本题以 CMU-MOSEI 英文多模态情感数据集为基础，围绕多模态特征提取与时序对齐、模态缺失条件下的鲁棒预测、可解释性情感预测三个层级开展系统性建模研究。

二、数据说明

本题提供如下数据与材料：

1. 附件 1：原始多模态样本

包含从 CMU-MOSEI 数据集中筛选的 100 条英文原始视频，具体描述如下：

- 附件包含 37 个子文件夹，其命名为原始视频标识 `video_id`；每个子文件夹包含一个或多个视频片段，命名 `clip_id.mp4`。`video_id` 和 `clip_id` 共同确定一条视频样本。所有视频时长介于 2.648 秒和 34.567 秒之间。
- 附件同时提供视频标注对应表（`label-100.xlsx`），每一行对应一个视频样本。其中，`video_id` 列和 `clip_id` 列对应子文件夹名称和文件名；`text` 列为该视频片段对应的英文转写文本；`label` 列为连续情感强度标注，取值： $[-3,0)$ （负向）、 0 （中性）、 $(0,3]$ （正向），数值越接近 -3 表示负向程度越强，越接近 3 表示正向程度越强； 0 仅属于中性，不同时归入负向或正向。；`annotation` 列为情感极性的英文文字标签，取值：Negative（负向情感）、Neutral（中性情感）、Positive（正向情感）。

注：本附件所有内容均由赛题统一提供。参赛队不得自行替换、增补、删除或修改样本及其标签；如需在模型处理中对极短、无效或异常片段采取掩码、截断或其他处理，应保留原始样本，并在报告中说明处理规则。

2. 附件 2：CMU-MOSEI 标准化多模态特征文件（.pkl 格式）

本附件提供：对 CMU-MOSEI 公开数据中有效的部分视频片段进行处理后、可直接用于建模的特征文件，包含对齐和非对齐各 4850 条样本，以及样本标签对应的 excel 表格。样本按训练/验证/测试划分组织，包括文本、语音、视觉三模态各自的特征矩阵、标签映射、维度定义与存储格式说明（数据形式和解释见后面补充说明）。参赛队可以根据情况选择附件 2 中的数据开展问题 2 和问题 3 建模。

3. 附件 3：模态局部缺失专项测试集

包含若干无标签样本对应的处理后的带有随机模态缺失的特征文件，用于模态局部缺失专项测试。其中，模态缺失指样本局部信息不可用，即一个或多个模态中存在特征值全部为零的随机连续序列区间。

4. 附件 4：可解释性专项测试集

包含若干无标签真实场景视频及对应的三模态特征文件，三模态信息完整。特征字段、张量维度和时序组织方式与附件 2 保持一致；配套原始视频用于将模型给出的关键证据时间位置回看至文本、语音或视觉片段。

注：以上特征文件分别提供与附件 2 aligned_50.pkl 和 unaligned_50.pkl 对应的两套版本。参赛队应在训练、验证和专项测试中保持同一特征版本和同一输入接口。

三、需要解决的问题

本题依次要求完成多模态特征提取与时序对齐、模态缺失条件下的鲁棒情感预测、可解释性情感预测三项建模任务。

问题 1：多模态情感特征提取与时序对齐的数学建模

基于附件 1 提供的原始视频，建立文本、语音、视觉三模态情感特征提取与时序对齐的数学模型，模型须明确原始样本的处理流程、各模态情感特征的定义与提取方法、跨模态时序对齐的规则与实现逻辑。参赛队可使用开源工具或预训练模型完成特征提取基础处理。

参赛队在问题 1 中自行生成的特征文件需达到如下要求：一是原始样本覆盖完整性，包括样本编号、模态文件与输出特征是否一一对应；二是时序组织的可核验性，包括序列位置、有效长度、填充规则及其与原始素材的对应记录；三是方法的合理性与可复现性，包括特征提取工具、关键参数、处理日志和复现实验说明。

问题 2：模态信息缺失条件下的情感预测建模与验证

针对局部时段模态信息缺失的复杂场景，构建鲁棒性多模态情感预测模型。模态局部缺失指文本、语音或视觉中的部分连续时间段不可用，并不表示整个模态完全不存在。模型需在局部模态信息不完整的条件下稳定输出情感极性与情感强度预测，并分析缺失模态类型、缺失位置、缺失时长对预测性能的影响规律。

使用所建立的鲁棒性情感预测模型，对附件 3 中的测试样本完成推理预测，按第四条结果与提交说明提交预测结果文件。

问题 3：可解释性情感预测建模与验证

针对三模态信息完整但决策依据难以追溯的场景，构建可解释性多模态情感预测模型。可解释性指模型在给出情感预测结果的同时，能够以可量化、可复核的方式说明本次判断主要依据了哪些输入信息，包括不同模态在当前样本中的作用差异、对预测起主要作用的模态，以及各模态中与判断密切相关的局部片段。

模型需同时输出情感极性、情感强度预测结果，以及主要参考模态、模态作用程度、关键证据定位等可解释性分析结果，关键证据需可对应至原始文本片段、语音时段或视觉关键帧。

使用所建立的可解释性情感预测模型，对附件 4 中的测试样本完成推理预测，按第四条要求格式提交预测结果与解释文件。

说明：问题 2 与问题 3 的模型均在附件 2 提供的训练集上学习模型参数，在验证集上选择模型结构、超参数与决策阈值；附件 3、附件 4 为无标签专项测试集，仅用于相应专项模型的最终推理预测。分类任务（情感极性）采用准确率（Accuracy）、F1 值评价，回归任务（情感强度）采用平均绝对误差（MAE）、皮尔逊相关系数评价。

四、结果与提交说明

所有提交内容需严格对应本题各问题的任务要求，在参赛论文中完整、规范呈现。

（一）论文正文要求

参赛论文需按照竞赛统一格式规范撰写，所有问题的方法原理、建模过程、实验结果与分析结论均须系统纳入正文，不得仅存放于附件。

1.论文组织要求

论文应依次回应问题 1、问题 2 和问题 3，并在每个问题中说明数据输入、方法或模型、求解过程、评价指标和主要结果。参赛队可自主安排理论推导、算法细节、消融实验和创新讨论，但不得省略题目规定的核心输出、结果文件说明和必要的实验依据。

2.问题 1 相关内容

(1)特征提取与时序对齐整体方案：说明文本、语音、视觉三类模态的特征定义、提取方法与所用工具，阐述时序对齐的核心规则与跨模态时间对应逻辑，完整呈现从原始素材到标准化时序特征的处理流程。

(2)特征文件规范与全量结果汇总：说明自生成特征文件的存储格式、组织结构与读取方式；以表格形式汇总附件 1 全部 100 条样本的特征提取结果，包含样本编号、模态类型、原始有效时长、特征维度、对齐粒度等关键信息，确保全量结果可追溯、可核验。全量原始特征文件随附件提交。

(3)典型样本验证：选取至少 1 个典型样本，展示其文本片段、对应语音时段、视频帧段与三类特征的对应关系，呈现时序对齐效果。

(4)方案可复现性说明：标注所用工具与模型的版本、核心参数及完整运行流程，保障结果可复现。

3.问题 2 相关内容

(1)鲁棒性模型的建模原理、网络结构、目标函数、训练方案与关键参数；

(2)缺失模态类型、缺失率对预测性能影响的规律分析与消融实验结果；

(3)附件 3 测试集的全量预测结果汇总与结果展示；

(4)验证集上的基础性能评价、可视化分析与错误归因结论。

4.问题 3 相关内容

- (1)可解释性模型的建模原理、网络结构、目标函数、训练方案与关键参数；
- (2)典型样本解释卡（含预测结果、三模态作用程度、主要参考模态及关键证据定位）；
- (3)主要参考模态内局部片段重要性分布可视化、三模态作用差异对比分析；
- (4)附件 4 测试集的全量预测与解释结果汇总；
- (5)验证集上的基础性能评价、可视化分析与错误归因结论。

（二）附件提交要求

1.可复现核心材料：问题 1 自主生成的 100 条样本多模态时序特征文件，问题 2 和问题 3 两类专项模型的核心代码、说明文档、模型参数文件、配置文件与运行环境说明。代码须附带详细运行说明、数据集处理规则与参数配置，确保实验结果可完整复现。

2.专项测试结果文件：附件 3 测试集预测结果 CSV 文件、附件 4 测试集预测与解释结果 CSV 文件，文件命名清晰规范。

总附件大小须符合竞赛统一要求（ $\leq 50\text{MB}$ ），所有提交材料中严禁出现参赛单位、队员姓名、队伍编号等身份信息，违者按竞赛违规处理。

五、补充说明

1.数据与工具使用规范

本题所有建模任务均以赛题提供的 CMU-MOSEI 系列数据为唯一数据来源，参赛队不得引入其他任何公开或私有情感数据集参与模型训练、微调、参数优化、阈值选择或结果统计。

参赛队可在合理论证前提下，对附件 2 的训练集、验证集数据开展清洗、重采样、标准化等基础预处理，也可基于该数据开展消融实验与规律分析。

本题所有问题的建模工作均允许使用开源特征提取工具、预训练模型与开源算法框架，但须在论文中统一标注所涉工具与模型的名称、版本号及核心参数设置，明确其使用场景。

2.关于附件 2 的数据解释和样例说明

CMU-MOSEI 数据集提供两个已处理特征文件：aligned_50.pkl 和 unaligned_50.pkl，选手可自行选择使用，但须在训练、验证和专项测试中保持同一版本。二者均以 Python Pickle 字典格式保存。

文件结构：最外层按数据用途划分为`train`、`valid`、`test`三个顶层键。进入任一顶层键后，文件按字段集中保存该划分下的全部样本，即`data['train']['audio']`表示训练集所有样本的语音特征。各字段中同一索引位置的取值共同描述同一条样本，读取方式为`data['train'][j][字段名]`（`j`为样本索引，Python 从 0 开始计数），而非`data['train'][j][字段名]`。

特征形状与时序组织：两个文件的矩阵形状如下表所示，第一维`N`为当前划分的样本数，后两维分别为序列位置数和特征维度。

表 1 附件 2 两类特征文件的矩阵形状

文件	文本特征	语音特征	视觉特征	时序含义
aligned_50.pkl	(N, 50, 768)	(N, 50, 74)	(N, 50, 35)	三模态均按最多 50 个相互对应的序列位置组织
unaligned_50.pkl	(N, 50, 768)	(N, 500, 74)	(N, 500, 35)	文本保持 50 个位置；语音、视觉保留独立时序，有效长度见 audio_lengths、vision_lengths。

文件名中的 50 表示统一采用的最大序列位置数。所有样本存储形状相同，有效位置数可能小于最大长度，其余位置按文件规则填充。语音特征维度 74 由原始提取流程确定，本赛题将其视为统一输入维度，不对其每一维的具体物理含义作额外假定。

字段含义：主要字段如表 2 所示。

表 2 附件 2 主要字段含义

字段	含义	使用说明
id	字符串，格式为 video_id\$_\$clip_id	唯一标识一条建模样本
raw_text	原始英文转写文本	用于阅读核查，非数值特征
text	连续文本特征，50×768	已经预计算的 BERT 类文本表示，可直接输入模型

字段	含义	使用说明
text_bert	BERT 三路整数输入， 3×50	词元编号、注意力掩码和分段编号，需经 BERT 编码后使用；与 text 二选一，不构成第四种模态
audio / vision	语音或视觉时序特征	形状见表 1
annotations / classification_labels / regression_labels	文字极性、三分类标签、连续情感强度	情感标注定义与问题 2 一致

读取示例：

```
split = data['train']
j = 1
sample_id = split['id'][j]           # 样本标识
audio = split['audio'][j]            # 语音特征
vision = split['vision'][j]           # 视觉特征
text = split['text'][j]               # 文本特征
label = split['regression_labels'][j] # 情感强度标签
```

附件 2 保存的是处理后的数值特征和标签，不保存原始视频帧、音频波形或逐词时间戳。为帮助参赛者理解这些字段在原始视频中的来源及其时间关系，图 3 用一个独立的多模态样本示意同一时间轴上的视觉、音频、文本和标签信息。

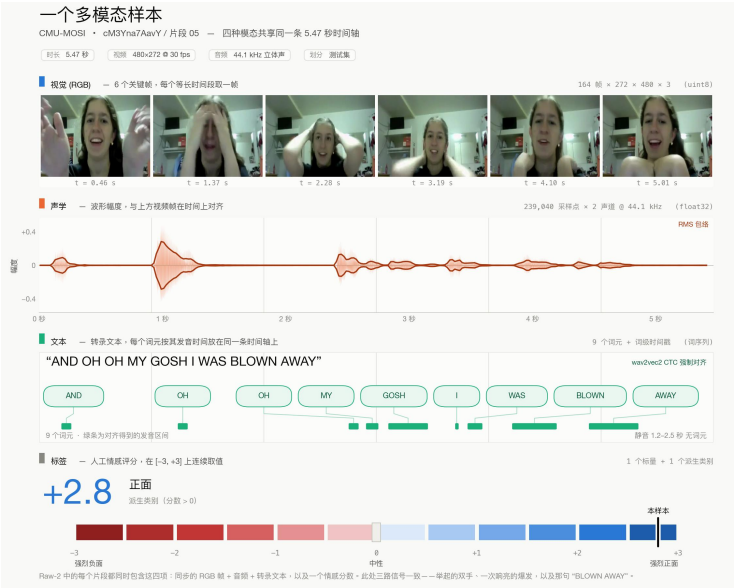


图 1 多模态视频片段的原始信号、文本时间戳与情感标签示例

参考文献

- [1] Zhang S, Yang Y, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects[J]. Expert Systems with Applications, 2024, 237: 121692.
- [2] Qiu X, Fang Y, Zhou Q, et al. Beyond Missing Modalities: Hypergraph Conditioned Diffusion for Uncertainty-Aware Multimodal Emotion Recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2026.
- [3] Yang Z, Li M. Factorize, Reconstruct, Enhance: A Unified Framework for Multimodal Sentiment Analysis[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2026.
- [4] Zhuang Y, Liu M, Bai W, et al. CMAD: Correlation-Aware and Modalities-Aware Distillation for Multimodal Sentiment Analysis with Missing Modalities[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2025.
- [5] Zhu A, Hu M, Wang X, et al. Proxy-Driven Robust Multimodal Sentiment Analysis with Incomplete Data[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). 2025.
- [6] Mai S, Han S. Learning Invariant Modality Representation for Robust Multimodal Learning from a Causal Inference Perspective[C]//Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL). 2026.
- [7] Fang Y, Huang W, Wan G, et al. EMOE: Modality-Specific Enhanced Dynamic Emotion Experts[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2025.
- [8] Wan J, Gong C, Fu G. Locate and Explain: Joint Multimodal Emotion Cause Extraction and Summarization in Conversation[C]//Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL). 2026.
- [9] Mai S, Han S. CaReFlow: Cyclic Adaptive Rectified Flow for Multimodal Fusion[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2026.
- [10] 王楠, 王淇, 欧阳丹彤. 基于知识蒸馏与动态调整机制的多模态情感分析模型[J]. 计算机学报, 2025, 48 (8): 1923-1942.