自动化学报
ACTA AUTOMATICA SINICA

基于视觉语言模型的多模态学生参与度预测方法

沈双宏 秦子轩 苏喻 王士进 黄振亚 孙登第

Multimodal Student Engagement Prediction Method Based on Vision-language Models

SHEN Shuang-Hong, QIN Zi-Xuan, SU Yu, WANG Shi-Jin, HUANG Zhen-Ya, SUN Deng-Di

基于视觉语言模型的多模态学生参与度预测方法

沈双宏[1] 秦子轩[2] 苏 喻[3,4] 王士进[5] 黄振亚[6] 孙登第[2]

摘 要 随着在线教育的普及, 学生参与度预测 (SEP) 已成为评估教学效能的核心任务. 尽管视觉语言模型 (VLMs) 在通用多模态表征学习中表现卓越, 但直接迁移至 SEP 领域时, 受限于对细粒度面部微表情及特定教学语境下宏观情绪的感知瓶颈, 难以实现视觉特征与高层语义标签的精准对齐. 为此, 提出基于 VLMs 的多模态学生参与度预测方法 VLM-SEP. 该方法先进行多层次参与度特征解耦, 从面部表情与动作姿态中提取结构化参与度特征; 再引入知识引导的注意力集中度推理, 建立离散视觉特征与参与度状态语言描述的显式映射; 最后通过跨模态融合决策, 结合视觉与文本信息实现参与度精准判别. 在三个公开数据集上的实验结果表明, 该方法有效提升了 VLMs 在 SEP 领域的适配性, 为学生参与度预测提供可解释解决方案.

关键词 视觉语言模型; 学生参与度预测; 多模态融合; 交叉注意力

引用格式 沈双宏, 秦子轩, 苏喻, 王士进, 黄振亚, 孙登第. 基于视觉语言模型的多模态学生参与度预测方法. 自动化学报, xxxx, **xx**(x): x−xx

DOI 10.16383/j.aas.c260173 **CSTR** 10.16383/j.aas.c260173

Multimodal Student Engagement Prediction Method Based on Vision-language Models

SHEN Shuang-Hong[1] QIN Zi-Xuan[2] SU Yu[3,4] WANG Shi-Jin[5] HUANG Zhen-Ya[6] SUN Deng-Di[2]

Abstract With the widespread adoption of online education, student engagement prediction (SEP) has emerged as a core task in evaluating teaching effectiveness. Although vision-language models (VLMs) have demonstrated exceptional performance in general multimodal representation learning, their direct transfer to the SEP domain is hindered by perceptual bottlenecks regarding fine-grained facial micro-expressions and macro-emotions within specific pedagogical contexts. Consequently, it remains challenging to achieve a precise alignment between visual features and high-level semantic labels. To address this issue, we propose VLM-SEP, a multimodal student engagement prediction method based on VLMs. Specifically, this method first performs multi-level engagement feature decoupling to extract structured engagement features from facial expressions and body postures; Subsequently, it introduces knowledge-guided attention concentration reasoning to establish an explicit mapping between discrete visual features and linguistic descriptions of engagement states; Finally, through cross-modal fusion decision-making, it integrates visual and textual information to achieve accurate engagement assessment. Experimental results on three public datasets demonstrate that the proposed method effectively enhances the adaptability of VLMs in the SEP domain, offering an interpretable solution for student engagement prediction.

Keywords vision-language models; student engagement prediction; multimodal fusion; cross-attention

Citation Shen Shuang-Hong, Qin Zi-Xuan, Su Yu, Wang Shi-Jin, Huang Zhen-Ya, Sun Deng-Di. Multimodal student engagement prediction method based on vision-language models. *Acta Automatica Sinica*, xxxx, **xx**(x): x−xx

在教育数字化转型的背景下, 在线教育平台的规模化部署促使教学模式向个性化、数据驱动范式

收稿日期 2026-03-09 录用日期 2026-05-07
Manuscript received March 9, 2026; accepted May 7, 2026
国家自然科学基金 (62507010, 62577019), 中国博士后科学基金 (2024M760725), 安徽省自然科学基金 (2408085QF212), 安徽省高校自然科学研究重大项目 (2025AHGXZK20027), 安徽省高校质量工程教学研究重大项目 (2023jujxwt006), 华南师范大学广东省心理健康与认知科学重点实验室开放课题, 安徽省高校数字化转型项目资助
Supported by National Natural Science Foundation of China (62507010, 62577019), China Postdoctoral Science Foundation (2024M760725), Anhui Provincial Natural Science Foundation (2408085QF212), Key Project of Natural Science Research in Anhui Provincial Institutions of Higher Education (2025AHGXZK20027), Key Teaching Research Project of the Quality Engineering Program in Anhui Provincial Institutions of Higher Education (2023jujxwt006), Opening Project of Key Laboratory of Mental Health and Cognitive Science (South China Normal University), and Anhui Provincial University Digital Transformation Project
本文责任编委 黄华
Recommended by Associate Editor HUANG Hua

1. 合肥综合性国家科学中心人工智能研究院 合肥 230011 2. 安徽大学人工智能学院 合肥 230601 3. 合肥师范学院计算机与人工智能学院 合肥 230061 4. 华南师范大学广东省心理健康与认知科学重点实验室 广州 510631 5. 科大讯飞股份有限公司 合肥 230088 6. 中国科学技术大学计算机科学与技术学院 合肥 230026
1. Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei 230011 2. School of Artificial Intelligence, Anhui University, Hefei 230601 3. School of Computer Science and Artificial Intelligence, Hefei Normal University, Hefei 230061 4. Guangdong Key Laboratory of Mental Health and Cognitive Science, South China Normal University, Guangzhou 510631 5. IFLYTEK Co., Ltd., Hefei 230088 6. School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026

演进[1−2]. 作为评估教学效能与构建个性化反馈闭环的核心任务, 学生参与度预测 (student engagement prediction, SEP) 的重要性日益凸显. 已有研究表明, 学生参与度不仅显著影响学习成效与知识保持率, 更是预测辍学风险的关键指标[3−4]. 随着人工智能技术与教育领域的深度融合, 自动化 SEP 系统为解决大规模在线教育中教师无法实时全量监测学生状态的难题提供可行路径. 然而, 实现稳健的 SEP 仍面临重大挑战: 如何在非受控的在线学习环境中, 从有限且含噪的视觉信号中准确提取特征, 并映射至抽象的多层次参与度标签[5−6].

相较于传统线下教学, 在线环境下的 SEP 面临数据异构性强、特征非结构化及跨场景泛化困难等多重挑战[7−9]. 现有方法主要可归纳为三类: 基于传统视觉特征的端到端学习方法、基于心理生理信号的多模态融合方法, 以及基于大规模预训练模型迁移的方法. 第一类方法依赖手工设计特征 (如眼部凝视、头部姿态) 或深度卷积神经网络自动提取特征进行直接分类, 但在原理上存在语义断层问题: 低层像素特征难以自然映射至高层参与度概念, 且无法有效捕捉参与度的动态演变; 第二类方法引入眼动追踪、脑电图 (electroencephalogram, EEG) 等生理信号, 虽在理论上信息量丰富, 但高部署成本与强侵入性严重制约其在真实教育场景中的普适性; 第三类方法利用预训练视觉–语言模型 (vision-language models, VLMs), 如 CLIP (contrastive language-image pre-training)、BLIP (bootstrapping language-image pre-training) 进行迁移学习, 尽管在通用视觉任务中表现卓越, 但直接应用于 SEP 时面临领域鸿沟: 由于模型缺乏对教学场景特定参与度表征 (如专注时的微表情、困惑时的眉间皱起) 的细粒度感知能力, 其应用效果受限.

综合来看, 参与度预测的核心挑战并非单纯的视觉识别, 而是深层的语义理解问题. 人类教师在判定参与度时, 并非机械地匹配面部特征, 而是基于表情变化、姿态调整等线索进行情境化推理[5]. 例如, 将“眉间皱起伴随瞳孔收缩”解读为困惑, 进而推断学生处于高认知负荷下的高参与状态. 这一推理过程构建从视觉观察到文本理解再到抽象概念生成的映射链条. 受认知心理学双重编码理论与情境认知理论的启发, 本文提出假设: 通过将视觉特征转化为文本描述, 并利用预训练 VLMs 强大的语义推理能力, 可实现更具解释性的参与度预测.

鉴于此, 本文提出一种融合视觉感知与显式规则推理的多模态框架——VLM-SEP (vision-language model for student engagement prediction), 旨在提升通用视觉语言模型在学生参与度预测任务中的表现. 在理论层面, 本文构建“视觉特征 → 语言描述 → 参与度概念”的分层映射机制, 并创新性地将教育心理学规则作为归纳偏置引入模型. 在方法论层面, VLM-SEP 采用递进式的三阶段架构: 首先是多层次参与度特征解耦, 通过多层次特征提取机制, 精准捕获面部表情与动作姿态等细粒度视觉线索. 其次是知识引导的注意力推理, 将教育心理学知识先验显式编码为推理规则, 借由思维链技术建立特征到注意力集中度的因果映射关系; 将前述视觉特征转化为自然语言的参与度描述. 最后是跨模态融合决策框架, 通过协同优化视觉与文本表征, 实现对学生参与度状态的精准判别. 实验表明, 该方法有效提升了视觉语言模型在 DAISEE、Emoti-W2023 以及 DIPSER 数据集上的预测精度, 并且具有出色的可解释性. 本文其余部分组织如下: 第 1 节对学生参与度预测任务进行形式化定义; 第 2 节回顾相关领域的现有工作; 第 3 节详细阐述 VLM-SEP 框架的网络架构与实现细节; 第 4 节开展广泛的对比实验并进行深入剖析; 第 5 节总结全文并对未来研究方向进行展望.

1 学生参与度预测任务

在教育数据挖掘与学习分析领域, 学生参与度预测旨在利用计算机视觉与模式识别技术, 自动量化学习者在教学交互过程中的认知与行为投入状态. 在本文关注的基于短时视频片段的应用场景中, 该任务可被形式化为受限时间窗口内的时空序列分类或有序回归问题. 其核心挑战在于: 如何构建有效的深度表征学习框架, 从蕴含面部表情、头部姿态、视线方向及肢体动作的动态视觉流中提取高区分度的时空特征, 并将其精确映射至预定义的参与度语义空间.

具体而言, 给定包含 N 个标注样本的数据集 $\mathcal{D} = \{(X_i,\, y_i)\}_{i=1}^{N}$. 对于第 i 个样本, 输入 X_i 为记录学习者状态的短视频片段. 按固定帧率采样后, 该片段可表示为由 T 帧图像构成的四维张量:

$$X_i = \{x_{i,\,1},\, x_{i,\,2},\, \cdots,\, x_{i,\,T}\} \in \mathbf{R}^{T\times C\times H\times W} \tag{1}$$

其中, T 为时间步长 (即序列帧数); C 为通道数 (对于 RGB 图像, $C = 3$); H 与 W 分别代表单帧图像的空间高度与宽度; 输入 X_i 对应的真实参与度标签记为 y_i. 若将参与度划分为 K 个离散等级, 则标签空间可定义为:

$$\mathcal{Y} = \{1,\, 2,\, \cdots,\, K\}, \quad y_i \in \mathcal{Y} \tag{2}$$

值得注意的是, 由于参与度等级间存在固有的

相对强弱与递进属性, 标签空间 $\mathcal{Y}$ 本质上呈现有序关系. 在此设定下, 模型训练的核心目标是学习一个参数化映射函数 f_θ, 该函数旨在捕获输入视频流的时空动态特征, 并输出对应的类别概率分布:

$$f_\theta : \mathbf{R}^{T\times C\times H\times W} \to \mathbf{R}^K \tag{3}$$

$$\hat{p}_i = \mathrm{Softmax}(f_\theta(X_i)) \tag{4}$$

其中, θ 为网络的可学习权重参数; $\hat{p}_i \in \mathbf{R}^K$ 表示样本 i 隶属于各参与度等级的预测概率向量.

2 相关工作

学生参与度预测研究经历从静态特征分析向时空动态建模的演进. 早期方法主要依赖如 ResNet (residual network) 的卷积神经网络提取帧级静态特征[10−12]. 为捕获时间维度的动态演变, Zhang 等[13] 提出 WSRGB-I3D 架构, 利用双流三维卷积网络 (three-dimensional convolutional neural network, 3D-CNN) 分别处理空间 RGB 数据与时间光流信息, 从而增强模型对光照变化的鲁棒性. 为进一步实现时空特征的解耦表征, Liao 等[14] 开发 SENet-LSTM 混合架构. 近年来, 得益于 Transformer 在长距离依赖建模方面的优势, Liu 等[15] 引入 Swin-Transformer, 有效平衡了局部细节与全局语义的表征能力. 在细粒度特征分析方面也取得显著进展: Huang 等[16] 提出 DERN 模型, 结合时间卷积网络 (temporal convolutional network, TCN) 与双向长短期记忆 (bidirectional long short-term memory, BiLSTM) 网络, 处理由 OpenFace 提取的视线与唇部几何特征序列[17]; Vedernikov 等[18] 则通过 TCCT-Net 将面部特征点序列转化为二维张量并进行小波变换, 大幅优化计算效率. 在多模态融合领域, Deng 等[19] 设计 MGAFR 框架, 引入图神经网络 (graph neural network, GNN) 以插补缺失模态; Singh 等[20] 提出的 VisioPhysioENet 则通过集成血容量脉搏 (blood volume pulse, BVP) 信号, 探究生理唤醒与显性行为之间的内在关联; Boitel 等[21] 提出 Mist 方法, 通过将学生静止状态和动作状态等多方面特征进行权重加和, 探究动静之间的关联对参与度变化的影响.

尽管上述监督学习方法在特定数据集上取得显著进展, 但受限于数据集固有的分布偏差, 导致其在跨域应用中的泛化性能严重受限. 相比之下, 基于海量数据预训练的视觉语言模型展现出强大的零样本推理能力. 然而, 将 VLMs 直接应用于 SEP 任务仍面临多重局限: 首先, 通用模型缺乏对特定教学情境下细微面部线索的敏感性; 其次, 全参数微调不仅计算成本极高, 且极易对领域数据中的噪声产生过拟合, 进而引发灾难性遗忘; 此外, 现有方法常忽略视觉与语言表征空间之间的语义鸿沟, 导致模型易产生幻觉. 针对这些问题, 研究人员展开广泛探索: Liu 等[22] 引入提示工程以约束预测空间; Zou 等[23] 利用多智能体辩论机制提升推理准确率; Qi 等[24] 则尝试对 VLMs 进行微调以适配下游任务. 尽管如此, 如何高效引导通用 VLMs 适配 SEP 等垂直领域任务, 同时充分挖掘并利用多模态间的互补性, 仍是当前亟待突破的关键挑战.

针对上述局限性, 本文提出 VLM-SEP 方法. 该方法旨在结合教育心理学先验知识, 通过多模态特征的精细化处理与多模态推理融合, 实现通用视觉语言模型在 SEP 任务中的高效迁移与应用.

3 基于视觉语言模型的多模态学生参与度预测方法

针对通用视觉语言模型在课堂参与度评估任务中存在的跨域语义鸿沟问题, 本文提出一种基于视觉语言模型的多模态学生参与度预测方法 (VLM-SEP), 如图 1 所示. 首先, VLM-SEP 对连续视频流进行多层次参与度特征解耦, 将视觉特征映射至面部表情与动作姿态两个正交的语义子空间, 旨在剥离教室背景、光影等非任务相关噪声, 确保模型能够精准锚定视线偏移、眼睑开合等细粒度参与度信号. 其次, 引入教育心理学先验知识引导 VLMs 进行注意力集中度推理, 将离散视觉特征转化为结构化自然语言描述; 通过将 Fredricks 参与度三维理论[25−26] 编码为提示词逻辑约束, 该模块实现从底层像素观测到高阶语义表征的跨维度映射, 为后续决策提供可解释的符号化输入. 最后, 提出一种跨模态交叉注意力融合框架, 深度协同视觉直观表征与文本逻辑推理信息, 利用双向交叉注意力建立起视觉感官特征与文本逻辑特征的动态关联, 输出高置信度的参与度级别判定结果.

3.1 多层次参与度特征解耦

在学生参与度预测任务中, 由于拍摄角度和遮挡等问题, 输入的原始视频数据通常伴随着显著的噪声干扰与高维特征冗余, 导致原始像素空间与高层参与度语义空间之间存在巨大的语义差异. 为弥合这一差异并实现信息的有效解耦, 本文依据教育心理学中“面部表情映射认知情感、动作姿态反映行为投入”的相关理论[27], 提出一种多层次参与度特征解耦模块, 旨在对连续视频流中的学生参与状态进行解耦分析与结构化重构, 具体提示词 P_{dec} 如

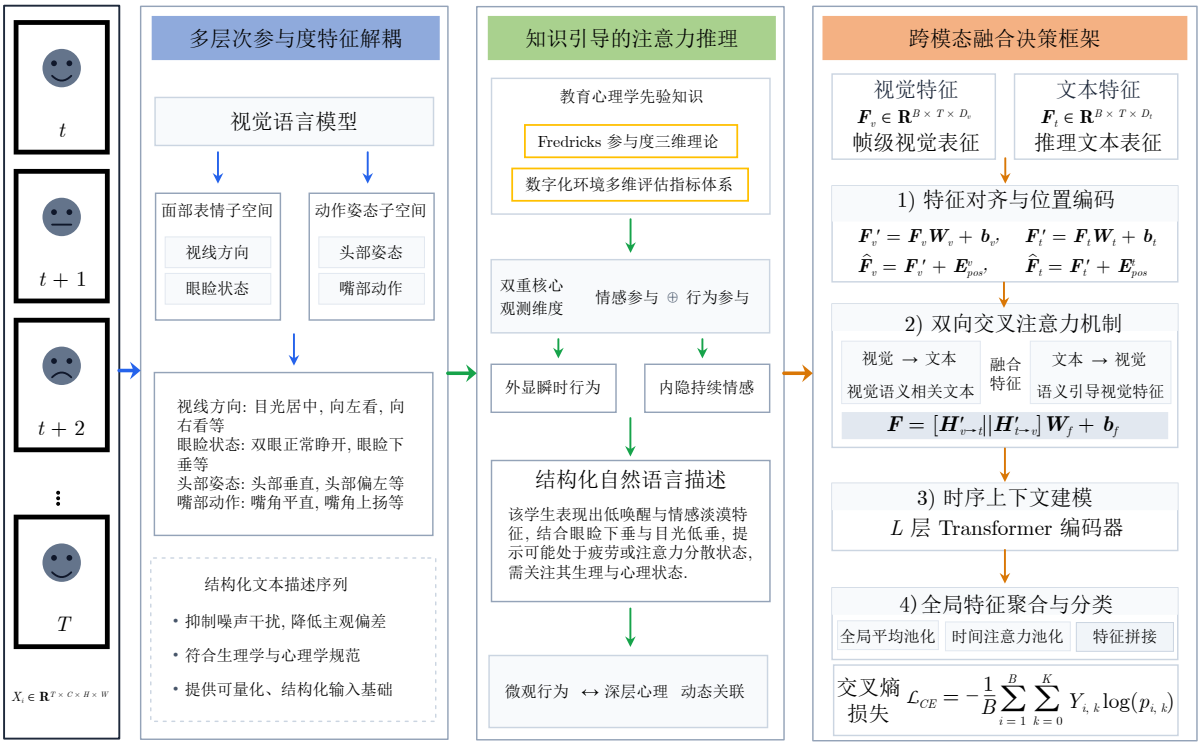


图 1 VLM-SEP 模型框架图

Fig. 1 The framework of the VLM-SEP model

图 2 所示.

为实现从高维像素空间到细粒度语义空间的映射, 本文定义解耦映射函数 $\Psi(\cdot)$. 给定输入视频序列 $X_i = \{x_{i,t}\}_{t=1}^{T}$, 该模块利用视觉语言模型 (例如, Qwen3-8B-VL[28]) 及解耦提示词 P_{dec} 将其转化为结构化的参与度特征文本描述序列 $s_{i,t}$. 特征解耦过程涵盖多个关键维度, 主要包括视线方向、眼睑状态、头部姿态及嘴部动作等.

$$s_{i,t} = \Psi(x_{i,t},\ P_{dec}) = \{a_1,\ a_2,\ \cdots,\ a_M\} \tag{5}$$

其中, a_i 表示图 1 中定义的特定语义属性 (如视线方向、眼睑状态等). 通过该映射, 原始视觉信号被转化为结构化的输入视觉描述 $S_i \in \mathbf{R}^{T \times D_t}$, 其中 D_t 为语义特征的维度, 从而抑制非任务相关的背景噪声. 该方法的核心优势在于, 通过将连续、高维且充满噪声的视觉信号, 离散化为符合生理学与心理学规范的标准语义描述, 不仅有效抑制遮挡等非关键信息的干扰, 最大限度地降低人工观察中固有的主观偏差, 更为后续的符号化推理分析提供客观、可量化且高度结构化的输入基础, 从而显著增强模型的可解释性与鲁棒性. 在特征提取的过程中, 本文设计提示词指令约束和验证特征类别函数用来保证提取结果高度符合指定内容, 避免生成规则以外的微表情特征, 确保后续时序建模阶段输入数据的高度一致性.

为验证 VLMs 特征提取的鲁棒性及幻觉抑制能力, 在格式稳定性层面, 本文在提示词中采用严格的 JSON Schema 约束与字段级枚举限定 (如视线方向仅允许 {“center”, “left”, “right”, “down”} 四个合法值), 配合输出端的正则表达式验证函数 $\mathcal{V}(\cdot)$ 对每一帧的返回结果进行格式校验. 在三个数据集共计 25 504 个样本上, 经过约束后提取到的结构化特征高度符合指定的特征结构和特征类别, 证明所提方法在格式输出上具有较高的稳定性. 其次, 在语义准确性层面, 本文随机抽取 15 个学生动作变化较为明显的样本共 240 帧图像进行人工检测, 其中包含情绪变化、动作变化、视线变化明显的样本. 结果表示, VLMs 提取的视线方向准确率为 81.3%, 眼睑状态准确率为 86.3%, 头部姿态准确率为 92.5%, 情绪状态准确率为 94.6%. 这些结果表明, VLMs 特征提取模块在实际应用中具有足够的可靠性.

3.2 知识引导的注意力推理

传统的自注意力机制作为纯数据驱动的“黑盒”模型, 在学生参与度预测任务中存在显著局限: 首先是噪声敏感性, 易受教室背景、光影波动等环境噪声干扰, 导致注意力权重偏离核心行为特征; 其

你是一名教育心理学领域的研究员, 专门通过观察学生的面部特征与姿态来分析课堂专注度.

请根据提供的目标视频帧 (清晰展示学生面部与头部姿态), 严格按照以下 5 个预设维度, 精准识别图像中学生的参与度特征. 具体要求如下:

从以下五个维度进行描述:

1) 视线方向: 从「目光居中、目光向上、目光向下、目光向左、目光向右、斜向上、斜向下」中选择.

2) 眼睑状态: 从「双眼正常睁开、微眯、眼睑下垂」中选择.

3) 头部姿态: 从「头部直立、身体前倾、头部侧倾、身体后仰、用手撑头」中选择.

4) 嘴部动作: 从「嘴角上扬、嘴角下撇、嘴角平直」中选择.

5) 情绪状态: 从「愉悦、困惑、悲伤、不耐烦、严肃、平淡」中选择.

请以 JSON 格式输出:

```
{
  "注视方向": "",
  "眼睑状态": "",
  "头部姿态": "",
  "嘴部动作": "",
  "情绪状态": ""
}
```

图 2 多层次参与度特征解耦模块的提示词

Fig. 2 Prompt for multi-level engagement feature decoupling module

次是语义断层, 由于缺乏先验知识约束, 低阶像素特征难以自发映射至高阶参与度概念, 存在逻辑失调风险; 最后是可解释性缺失, 其纯统计意义的权重分配往往违背教育心理学的判别准则. 针对上述挑战, 本文提出一种知识引导的注意力推理模块, 旨在通过引入教育心理学先验规则作为归纳偏置, 将离散的多维视觉特征映射至具备情境化逻辑的推理框架中. 该模块有效弥补了传统模型在逻辑推演与语义对齐上的不足, 实现从数据驱动到知识辅助决策的有效修正, 具体提示词如图 3 所示.

在理论构建层面, 本文借鉴文献 [25–26] 基于 Fredricks 参与度三维理论所构建的数字化环境多维评估指标体系. 考虑到单一视觉特征在表征认知参与时的固有局限性, 本文重点选取情感参与与行为参与作为测度注意力集中度的双重核心观测维度. 基于此, 本文在构建推理规则库时, 摒弃单一特征分析范式, 采用“外显瞬时行为–内隐持续情感”的协同建模策略: 一方面, 实现对视线焦点、头部姿态等显性行为特征的逻辑判别; 另一方面, 引入对情感语义类别 (如愉悦、悲伤、平静、专注及疲滞) 的细粒度识别, 并辅以针对状态突变等时序动态模式

你是一位深耕教育心理学与学习行为分析领域的资深专家研究员, 熟练掌握多模态学习分析技术与行为编码体系.

你的任务是基于 Heilporn 等[25] 提出的“参与度多维评估指标体系”(涵盖行为参与、情感参与、认知参与), 深度分析从视频中提取的关于某位学生面部与肢体状态的结构化观察记录. 你需要通过严密的逻辑推理, 评估该学生的注意力状态, 并给出专业的心理学诊断与归因.

输入数据格式:

```
{
  "注视方向": "",
  "眼睑状态": "",
  "头部姿态": "",
  "嘴部动作": "",
  "情绪状态": ""
}
```

请遵循以下教育心理学与行为特征映射逻辑进行推理:

1) 特征解码:
- 视线与头部姿态: 主要反映行为参与度与视觉注意力的定向.
- 眼睑状态: 主要反映生理唤醒水平, 是判断疲劳度或情感参与度的重要指标.
- 嘴部动作: 辅助反映认知参与度 (如默读代表深层加工) 或生理状态 (如打哈欠代表低唤醒).

2) 综合诊断: 将上述多维度特征结合, 判断特征之间是否一致 (如视线专注但眼睑半闭可能代表“强撑专注的疲劳状态”), 从而推导出核心的注意力状态.

3) 客观约束: 只能基于输入的特征进行推演, 绝对禁止过度解读或捏造输入中不存在的动作.

请严格按照以下 JSON 格式输出你的评估结果. 必须是合法的 JSON 字符串, 绝对不要包含任何多余的开头问候、过程解释或 Markdown 代码块标记 (如 JSON 标记)

{{
  "分析":"基于 Heilporn 等[25] 提出的参与度多维评估体系, 对该特征组合进行专业心理学诊断和归因分析, 字数严格控制在 100 字以内, 语言需学术、精准."
}}

图 3 知识引导的注意力推理模块的提示词

Fig. 3 Prompt for knowledge-guided attention reasoning module

的判定规则.

本文定义参与度三维理论 $\mathcal{K}$, 在具体技术实现层面, 本方法采用预训练视觉语言模型作为推理引擎, 利用大模型的上下文学习能力, 在推理提示词 P_{reason} 中显式注入指标体系的定义与判定准则, 从而强制模型在受限的语义空间内执行规范化的逻辑推演过程. 从数学层面看, 传统自注意力机制的权重分配可表示为 $\alpha_{ij} = \mathrm{Softmax}(q_i^{\mathrm{T}} k_j/\sqrt{d})$, 其中 q_i 是第 i 个查询向量, k_j 是第 j 个键向量, d 是特征维度, 注意力权重 α_{ij} 完全由数据驱动的查询–键匹配

决定, 缺乏对“哪些特征组合在教育学意义上构成参与度判据”的先验约束. 本文的知识引导机制通过在推理提示词 P_{reason} 中注入形式化规则 $\mathcal{R} = \{r_1, r_2, \cdots, r_M\}$ (例如, r_1:“视线居中 $\wedge$ 眼睑正常 $\rightarrow$ 行为参与度高”), M 表示样本最大数量, 将注意力分配从无约束的统计匹配转变为受知识约束的语义推理. 参与度三维理论 $\mathcal{K}$ 的构成包括三个层次: 1) 行为特征判定规则——基于 Fredricks 的行为参与维度, 定义视线、头部姿态等显性行为特征的分级判据; 2) 情感状态映射规则——基于情感参与维度, 建立面部表情与情感语义类别的对应关系; 3) 时序动态判定规则——定义状态突变等时序模式的识别准则. 上述知识以自然语言形式编码于 Prompt 中, 由 VLMs 的上下文学习能力实现从规则到推理的自动映射.

需要特别说明的是, 本方法并不依赖于 VLMs 的预训练语料覆盖 Heilporn 等[25] 提出的多维评估指标体系或 Fredricks 等[26] 提出的参与度三维理论. 相反, 本文的核心设计理念是将上述教育心理学知识显式编码为推理提示词 P_{reason} 中的形式化规则与判定准则 (如图 3 所示), 利用大语言模型强大的上下文学习能力, 使模型在推理时严格遵循注入的知识框架进行逻辑推演. 这一设计将模型定位为执行引擎, 理论知识由人类专家提供并编码为提示词约束, VLMs 负责在该约束空间内执行逻辑推理. 因此, 该方法对 VLMs 预训练语料的领域特异性要求极低, 具有良好的迁移性与可扩展性.

如图 3 所示, 推理模块通过知识引导映射函数 $\mathcal{F}_{\mathcal{K}}(\cdot)$ 将解耦后的输入视觉描述 S_i 转化为包含深度逻辑信息的文本特征序列 T_i. 在数学层面上, VLMs 作为执行引擎, 通过自回归条件生成方式对视觉特征与先验规则进行联合概率推演. 给定输入视觉描述 S_i 与知识引导提示词 P_{reason} (包含规则集合 $\mathcal{R}$), 模型生成长度为 ρ 的文本特征序列 $T_i = [w_1, w_2, \cdots, w_\rho]$ 的概率分布为:

$$P(T_i|S_i,\ P_{reason},\ \mathcal{R}) = \prod_{k=1}^{\rho} P_{vlm}\big(w_k|w_{<k},\ S_i,\ P_{reason}\big) \tag{6}$$

其中, P_{vlm} 表示大模型内化的联合概率分布; $w_{<k}$ 表示第 k 个词的生成依赖于前 $k-1$ 个词. 通过解码寻优, 模型输出概率最大的文本特征序列 T_i:

$$T_i = \arg\max_{\hat{T}} P(\hat{T}|S_i,\ P_{reason},\ \mathcal{R}) = \mathcal{F}_{\mathcal{K}}(S_i,\ P_{reason}) \tag{7}$$

最终生成的文本特征序列 $T_i = \{t_{i,1}, \cdots, t_{i,T}\} \in \mathbf{R}^{T\times D_t}$, 其中单帧文本表征 $t_{i,k}$ 是由 VLMs 隐层特征提取获得的 D_t 维向量. 这一机制不仅实现对视觉信号的符号化提炼, 更重要的是, 通过式 (6) 的条件生成路径, 确保推理过程严格遵循既定的教育学框架. 原本纯数据驱动的注意力分配机制由此被重塑为受先验知识强约束的白盒推理过程. 这一改进有效弥合通用大模型与特定垂直领域理论之间的鸿沟.

综上所述, 该推理模块深入剖析微观行为动作与深层心理状态之间的动态关联, 实现从底层离散视觉特征向高层结构化自然语言描述的语义映射, 确保量化结果兼具高数值精确性与符合教育心理学理据的逻辑一致性.

3.3 跨模态融合决策框架

为克服复杂教育场景下单一模态易失效及信息不对称的局限性, 本文在获取特定视频帧内学生注意力集中度的文本推理结果基础上, 引入原始视频帧图像, 构建一种跨模态融合决策框架. 该框架依托视觉语言模型, 旨在通过视觉直观表征与文本逻辑推理的高效协同, 实现多模态数据的特征嵌入与统一语义空间映射, 进而对学生课堂参与度进行高置信度判别.

具体而言, 为了对视频中的学生参与度进行精准分级, 本文提出一种基于双向交叉注意力机制的时空多模态融合网络, 旨在有效弥合异构模态间的语义鸿沟, 并捕捉视频序列中细粒度的时空依赖关系. 模型以帧级视觉特征 $\boldsymbol{F}_v \in \mathbf{R}^{B\times T\times D_v}$ 和文本特征 $\boldsymbol{F}_t \in \mathbf{R}^{B\times T\times D_t}$ 作为输入, 通过特征对齐与位置编码、双向跨模态交互、时序上下文建模、全局特征聚合四个递进阶段, 逐层提炼判别性表征, 最终输出学生参与度的分类预测. 其中, B 为批量大小, D_v 表示视觉特征的原始维度.

由于原始视觉特征与文本特征通常分布于异构的语义空间且维度不一致, 直接融合将引入噪声并阻碍跨模态信息交互. 本文首先采用可学习的线性投影层对双模态特征进行维度对齐与空间映射:

$$\boldsymbol{F}'_v = \boldsymbol{F}_v\boldsymbol{W}_v + \boldsymbol{b}_v, \quad \boldsymbol{F}'_t = \boldsymbol{F}_t\boldsymbol{W}_t + \boldsymbol{b}_t \tag{8}$$

其中, $\boldsymbol{W}_v \in \mathbf{R}^{D_v\times D}$、$\boldsymbol{W}_t \in \mathbf{R}^{D_t\times D}$ 为投影矩阵, D 为统一的隐层维度; $\boldsymbol{b}_v$、$\boldsymbol{b}_t$ 为偏置项.

考虑到学生参与行为具有显著的时间演化特性 (如注意力分散的动态过程), 为保留视频序列的时序依赖性, 本方法在对齐后的特征中注入可学习的时序位置编码:

$$\hat{\boldsymbol{F}}_v = \boldsymbol{F}'_v + \boldsymbol{E}^v_{pos}, \quad \hat{\boldsymbol{F}}_t = \boldsymbol{F}'_t + \boldsymbol{E}^t_{pos} \tag{9}$$

其中, $\boldsymbol{E}^v_{pos}$、$\boldsymbol{E}^t_{pos} \in \mathbf{R}^{T\times D}$ 为可学习的位置嵌入参数, 使模型能够自适应地编码帧间的时间先后关系, 而非依赖固定的正弦/余弦编码.

单一方向的跨模态注意力往往导致信息瓶颈, 即查询模态主导融合过程, 被查询模态的信息被选择性过滤. 为充分挖掘视觉与文本模态间的互补信息, 实现细粒度的语义对齐, 本文设计一个对称的双向交叉注意力模块. 该模块通过双向查询机制, 强制两种模态相互适应, 避免主导偏差. 具体包含两个对称的跨模态注意力分支: 一是视觉到文本的交叉注意力: 以视觉特征 $\hat{\boldsymbol{F}}_v$ 作为查询, 文本特征 $\hat{\boldsymbol{F}}_t$ 作为键和值, 通过注意力权重计算, 动态提取与当前视觉状态语义相关的文本上下文 (如学生的表情与对应的行为描述之间的关联):

$$\begin{cases} \boldsymbol{A}_{v\to t} = \mathrm{Softmax}\left(\dfrac{\hat{\boldsymbol{F}}_v \boldsymbol{W}_Q^v (\hat{\boldsymbol{F}}_t \boldsymbol{W}_K^t)^{\mathrm{T}}}{\sqrt{D_k}}\right) \\ \boldsymbol{H}_{v\to t} = \boldsymbol{A}_{v\to t}(\hat{\boldsymbol{F}}_t \boldsymbol{W}_V^t) \end{cases} \tag{10}$$

二是文本到视觉的交叉注意力: 以文本特征 $\hat{\boldsymbol{F}}_t$ 作为查询, 视觉特征 $\hat{\boldsymbol{F}}_v$ 作为键和值, 利用高层语义在视觉空间中进行特征寻址, 强化对关键视觉区域 (如面部表情、肢体动作) 的关注:

$$\begin{cases} \boldsymbol{A}_{t\to v} = \mathrm{Softmax}\left(\dfrac{\hat{\boldsymbol{F}}_t \boldsymbol{W}_Q^t (\hat{\boldsymbol{F}}_v \boldsymbol{W}_K^v)^{\mathrm{T}}}{\sqrt{D_k}}\right) \\ \boldsymbol{H}_{t\to v} = \boldsymbol{A}_{t\to v}(\hat{\boldsymbol{F}}_v \boldsymbol{W}_V^v) \end{cases} \tag{11}$$

其中, $\boldsymbol{W}_Q^v$、$\boldsymbol{W}_K^v$、$\boldsymbol{W}_V^v \in \mathbf{R}^{D\times D_k}$ 分别代表视觉特征的查询 (query)、键 (key) 和值 (value) 线性投影矩阵; $\boldsymbol{W}_Q^t$、$\boldsymbol{W}_K^t$、$\boldsymbol{W}_V^t \in \mathbf{R}^{D\times D_k}$ 分别代表文本特征的查询、键和值线性投影矩阵; D_k 为注意力头的特征维度. 上述过程采用多头注意力机制实现多子空间并行学习, 并结合残差连接与层归一化以稳定网络训练:

$$\begin{cases} \boldsymbol{H}'_{v\to t} = \mathrm{LN}(\hat{\boldsymbol{F}}_v + \mathrm{MHA}_v(\hat{\boldsymbol{F}}_v,\ \hat{\boldsymbol{F}}_t,\ \hat{\boldsymbol{F}}_t)) \\ \boldsymbol{H}'_{t\to v} = \mathrm{LN}(\hat{\boldsymbol{F}}_t + \mathrm{MHA}_t(\hat{\boldsymbol{F}}_t,\ \hat{\boldsymbol{F}}_v,\ \hat{\boldsymbol{F}}_v)) \end{cases} \tag{12}$$

其中, MHA_v、MHA_t 分别表示以视觉特征为查询、以文本特征为查询的多头注意力 (multi-head attention) 模块; LN 表示层归一化 (layer normalization) 操作.

随后, 将两个分支的输出在特征维度进行拼接, 并通过一个线性融合投影层 $\boldsymbol{W}_f \in \mathbf{R}^{2D\times D}$ 降维至 D, 得到蕴含双向语义关联的跨模态融合特征 $\boldsymbol{F} \in \mathbf{R}^{B\times T\times D}$:

$$\boldsymbol{F} = [\boldsymbol{H}'_{v\to t} \,||\, \boldsymbol{H}'_{t\to v}]\boldsymbol{W}_f + \boldsymbol{b}_f \tag{13}$$

其中, $\boldsymbol{b}_f$ 为该线性投影层的偏置项; $||$ 表示沿特征维度的拼接操作, 该操作确保视觉外观线索与文本语义线索在帧级别实现深度耦合.

在完成帧级别的多模态融合后, 需进一步建模视频序列的全局时空演变规律 (如参与度的渐变或突变). 模型采用 L 层 Transformer 编码器对融合特征 $\boldsymbol{F}$ 进行时序关系建模. 每层包含多头自注意力与前馈网络:

$$\boldsymbol{F}^{(l)} = \mathrm{FFN}^{(l)}\left(\mathrm{LN}\left(\boldsymbol{F}^{(l-1)} + \mathrm{MHSA}^{(l)}(\boldsymbol{F}^{(l-1)})\right)\right), \quad l = 1,\ \cdots,\ L \tag{14}$$

其中, $\boldsymbol{F}^{(0)} = \boldsymbol{F}$ 为初始输入特征序列; $\mathrm{MHSA}^{(l)}$ 表示第 l 层的多头自注意力模块, 用于捕获全局时间依赖关系; $\mathrm{FFN}^{(l)}$ 为第 l 层的前馈神经网络 (feed-forward network).

通过自注意力机制, 模型能够自适应地关注对判断参与度至关重要的关键帧片段 (如学生突然转头、表情变化等微行为), 捕获长距离依赖关系, 输出具有全局上下文感知的时序特征 $\boldsymbol{Z} \in \mathbf{R}^{B\times T\times D}$.

为获得视频级别的全局表征, 消除帧间冗余并保留最具判别性的信息, 模型对时序特征 $\boldsymbol{Z}$ 在时间维度上应用全局平均池化 (global average pooling, GAP) 操作:

$$\boldsymbol{z}_{global} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Z}_{:,\,t,\,:},\ \boldsymbol{Z}_{:,\,t,\,:} \in \mathbf{R}^{B\times D} \tag{15}$$

此外, 为增强模型对关键时间步的敏感度, 同时引入时间注意力池化作为互补, 通过学习时间权重 α_t 生成加权特征, 随后与全局平均池化特征拼接形成鲁棒的融合表征.

最终, 该全局表征 $\boldsymbol{z}_{global}$ 被送入由两层线性层、GELU 激活函数和 Dropout 正则化组成的多层感知机 (multi-layer perceptron, MLP) 分类头中, 映射到 K 个参与度级别 (如高、中、低参与度) 的逻辑值空间:

$$\boldsymbol{y} = \mathrm{MLP}(\boldsymbol{z}_{global}) = \mathrm{GELU}(\boldsymbol{z}_{global}\boldsymbol{W}_1 + \boldsymbol{b}_1)\boldsymbol{W}_2 + \boldsymbol{b}_2 \tag{16}$$

其中, $\boldsymbol{W}_1 \in \mathbf{R}^{D\times D_m}$ 与 $\boldsymbol{W}_2 \in \mathbf{R}^{D_m\times K}$ 分别为 MLP 中两个线性层的权重矩阵, D_m 为隐藏层维度; $\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 为相应的偏置项; GELU 表示高斯误差线性单元 (Gaussian error linear unit) 激活函数; $\boldsymbol{y} \in \mathbf{R}^{B\times K}$ 为模型针对不同参与度等级的预测输出, 随后通过 Softmax 函数转换为概率分布, 用于计算交叉熵损失并指导模型优化.

$$\begin{cases} p_{i,\,k} = \dfrac{\exp(y_{i,\,k})}{\sum\limits_{j=1}^{K}\exp(y_{i,\,j})} \\ \mathcal{L}_{CE} = -\dfrac{1}{B}\sum\limits_{i=1}^{B}\sum\limits_{k=1}^{K} Y_{i,\,k}\log(p_{i,\,k}) \end{cases} \tag{17}$$

其中, $y_{i,\,k}$ 表示模型对第 i 个样本属于第 k 类的逻辑输出值; $p_{i,\,k}$ 满足 $p_{i,\,k} \in (0,\,1)$ 且 $\sum_{k=1}^{K} p_{i,\,k} = 1$; $Y_{i,\,k}$ 为真实标签的独热编码, 如果第 i 个样本的真实类别为 k, 则 $Y_{i,\,k} = 1$, 否则 $Y_{i,\,k} = 0$.

VLM-SEP 的具体算法如算法 1 所示.

算法 1. VLM-SEP 算法介绍

输入. 原始视频序列 $\mathcal{X} = \{X_1,\, X_2,\, \cdots,\, X_B\}$, 其中 $X_i \in \mathbf{R}^{T\times C\times H\times W}$, 真实参与度标签 $Y \in \{0,\,1\}^{B\times K}$, 教育心理学先验规则库.

输出. 学生参与度预测概率分布 P, 模型交叉熵损失 $\mathcal{L}_{CE}$.

1) 初始化: 模型参数 $\boldsymbol{W}_v$、$\boldsymbol{W}_t$、$\boldsymbol{W}_f$, 位置编码 $\boldsymbol{E}^{v}_{pos}$、$\boldsymbol{E}^{t}_{pos}$, Transformer 层及分类头 MLP 参数.
2) **for** 每个批次 $X_i \in \mathcal{X}$ **do**
3) 提取离散视觉特征;
4) 将离散视觉特征转化为文本描述;
5) 获取帧级视觉特征 $\boldsymbol{F}_v \in \mathbf{R}^{B\times T\times D_v}$ 与文本特征 $\boldsymbol{F}_t \in \mathbf{R}^{B\times T\times D_t}$;
6) $\boldsymbol{F}'_v \leftarrow \boldsymbol{F}_v\boldsymbol{W}_v + \boldsymbol{b}_v$;
7) $\boldsymbol{F}'_t \leftarrow \boldsymbol{F}_t\boldsymbol{W}_t + \boldsymbol{b}_t$;
8) $\hat{\boldsymbol{F}}_v \leftarrow \boldsymbol{F}'_v + \boldsymbol{E}^{v}_{pos}$;
9) $\hat{\boldsymbol{F}}_t \leftarrow \boldsymbol{F}'_t + \boldsymbol{E}^{t}_{pos}$;
10) $\boldsymbol{H}'_{v\to t} \leftarrow \mathrm{LN}(\hat{\boldsymbol{F}}_v + \mathrm{MHA}_v(\mathrm{query} = \hat{\boldsymbol{F}}_v,\ \mathrm{key} = \hat{\boldsymbol{F}}_t,\ \mathrm{value} = \hat{\boldsymbol{F}}_t))$;
11) $\boldsymbol{H}'_{t\to v} \leftarrow \mathrm{LN}(\hat{\boldsymbol{F}}_t + \mathrm{MHA}_t(\mathrm{query} = \hat{\boldsymbol{F}}_t,\ \mathrm{key} = \hat{\boldsymbol{F}}_v,\ \mathrm{value} = \hat{\boldsymbol{F}}_v))$;
12) 拼接并融合: $\boldsymbol{F} \leftarrow [\boldsymbol{H}'_{v\to t} \,||\, \boldsymbol{H}'_{t\to v}]\boldsymbol{W}_f + \boldsymbol{b}_f$;
13) $\boldsymbol{F}^{(0)} \leftarrow \boldsymbol{F}$;
14) **for** $l = 1$ to L **do**
15) $\tilde{\boldsymbol{F}}^{(l)} \leftarrow \mathrm{LN}(\boldsymbol{F}^{(l-1)} + \mathrm{MHSA}^{(l)}(\boldsymbol{F}^{(l-1)}))$;
16) $\boldsymbol{F}^{(l)} \leftarrow \mathrm{FFN}^{(l)}\big(\tilde{\boldsymbol{F}}^{(l)}\big)$;
17) **end for**
18) $\boldsymbol{Z} \leftarrow \boldsymbol{F}^{(L)}$;
19) 全局平均池化: $\boldsymbol{z}_{global} \leftarrow \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{Z}_{:,\,t,\,:}$;
20) $\boldsymbol{y} \leftarrow \mathrm{MLP}(\boldsymbol{z}_{global})$;
21) 计算预测概率: $p_{i,\,k} \leftarrow \exp(y_{i,\,k})/\sum_{j=1}^{K}\exp(y_{i,\,j})$;
22) $\mathcal{L}_{CE} \leftarrow -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} Y_{i,\,k}\log(p_{i,\,k})$;
23) 更新网络参数并反向传播 $\mathcal{L}_{CE}$.
24) **end for**
25) **return** P、$\mathcal{L}_{CE}$

4 实验

为全面评估本文提出的基于视觉语言模型的多模态学生参与度预测方法 (VLM-SEP) 的有效性与泛化能力, 本节开展多维度的实验分析. 首先, 明确统一的软硬件实验环境, 并详细介绍采用的三个学生参与度公开基准数据集 (DAISEE[29]、EmotiW2023[30] 与 DIPSER[31]) 以及涵盖传统深度学习与前沿视觉语言大模型的对比基线. 随后, 通过广泛的对比实验定量验证该方法在预测精度与鲁棒性上的显著优势, 并利用消融实验深入解析多模态特征融合及双向交叉注意力等核心组件的具体贡献. 此外, 本节还进一步利用 t-SNE (t-distributed stochastic neighbor embedding) 算法对解耦前后的特征分布进行可视化, 验证多层次特征解耦模块抑制噪声的有效性; 通过视觉−文本交叉注意力热力图分析, 证明模型在跨模态语义映射上的精准对齐能力. 最后, 结合具体的学生参与度视频帧片段进行案例分析, 直观展示 VLM-SEP 模型基于教育心理学先验知识的推理过程, 验证其在真实教学场景中的高可解释性与应用价值.

4.1 实验环境

为确保实验结果的可复现性与计算性能的一致性, 本文所有实验均在统一的硬件与软件环境下执行. 硬件平台部署 Intel(R) Xeon(R) Platinum 8470Q 中央处理器 (25 vCPU) 以及 NVIDIA GeForce RTX 5090 图形处理单元 (graphics processing unit, GPU). 软件环境基于 Ubuntu 22.04 操作系统, 采用 PyTorch 深度学习框架构建并训练模型. 在视觉语言模型的部署与参数微调阶段, 集成 Transformers 库与 OpenCV 库的最新稳定版本. 为消除异构视频编码格式引入的域偏差, 所有输入视频在预处理阶段均经过时序重采样, 统一帧率标准为 30 fps, 实验中每个学生视频采用 16 帧的时序输入长度. 在模型构建与训练策略上, 本文采用分步优化的方案. 为了平衡大模型推理的计算负载, 特征提取与文本转化步骤在数据预处理阶段离线完成. 在跨模态融合决策网络的训练过程中, 预训练的视觉语言模型底座处于冻结状态, 不参与反向传播, 仅针对可学习的线性投影层、双向跨模态注意力模块以及后续的时序编码器进行端到端参数更新.

VLM-SEP 的分步优化策略分为两个阶段: 1) 离线预处理阶段: 第 3.1 节的多层次参与度特征解耦与第 3.2 节的知识引导的注意力推理均在数据预处理阶段离线完成. 在此阶段, 对所有视频样本逐帧调用 VLMs 提取特征并缓存. 该过程将 VLMs 的功能限定为纯粹的感知引擎, VLMs 权重完全冻结. 此设计基于两点理论考量: 首先是为规避灾难性遗忘. 学生参与度预测数据集规模通常远小于 VLMs

的预训练图文语料库, 若进行全参数端到端微调, 极易破坏大模型原本丰富的通用视觉多模态表征能力; 其次是受限于极高的计算开销以及显存瓶颈. 2) 在线训练阶段: 第 3.3 节的跨模态融合决策框架以预缓存的视觉与文本双模态时序序列作为输入进行端到端训练. 在该阶段中, 仅可学习的线性投影层 $\boldsymbol{W}_v$、$\boldsymbol{W}_t$、双向交叉注意力模块、时序 Transformer 编码器及 MLP 分类头计算梯度并更新, VLMs 底座始终保持冻结状态. 如算法 1 所示, 步骤 3) 和步骤 4) 对应离线特征提取, 步骤 5) 及以后为在线优化与融合. 这种解耦式训练策略不仅大幅提升训练效率, 且能基于同一套 VLMs 特征针对不同的下游交互模块进行灵活消融.

4.2　数据集与评估指标

为验证所提方法的有效性与泛化能力, 本文选取三个广泛应用于学生参与度评估的基准数据集: DAISEE、EmotiW2023 与 DIPSER. 数据分布情况如表 1 所示, 具体介绍如下:

表 1　数据集分布表

Table 1　Dataset distribution table

数据集	训练集	验证集	测试集	总计
DAISEE	5 466	1 712	1 708	8 886
EmotiW2023	5 443	1 427	1 074	7 944
DIPSER	5 560	1 470	1 644	8 674

● DAISEE 数据集. 该数据集包含来自 112 名受试者的 9 068 个多标签视频片段, 总时长约为 25 h, 涵盖四种情感状态的四级评分标注. 为确保与现有相关工作的可比性, 本文严格遵循标准的数据划分协议同时进行数据清洗及异常样本剔除.

● EmotiW2023 数据集. 该数据集专用于学生课堂参与度分析, 每个视频片段约为 10 s. 鉴于官方测试集标签未公开, 本文将原始训练集按受试者维度以约 5 : 1 的比例重新划分为训练集与测试集, 并严格控制各子集中类别分布的均衡性和剔除帧率异常及重采样后时长不达标的噪声样本.

● DIPSER 数据集. 该数据集是一个面向多模态心理状态识别的高质量公开数据集, 采用高分辨率视频帧 (帧率 25 fps), 覆盖不同年龄段、性别受试者的面部表情、眼部运动、头部姿态等行为特征, 其标签采用 5 等级划分, 由五个专家分别标注. 本实验采用标签取众数和中位数的方式进行标签评定, 将每个学生样本以 10 fps 为划分单位进行切割并剔除特征不足的样本后, 最终共分成三个等级: 极低参与、中等参与、高度参与.

为了全面评估模型的性能, 本文采用准确率 (accuracy)、F1 分数 (F1-score) 以及平均绝对误差 (mean absolute error, MAE) 作为核心评估指标. 具体而言, 准确率用于直观反映模型整体的预测正确比例; F1 分数兼顾精确率与召回率, 用于评估模型在正负样本上的综合分类效能; 平均绝对误差则用于量化预测等级与真实标签之间的平均偏离程度.

4.3　对比模型

本文选取近年来传统深度学习方法和基于视觉语言模型的方法作为基线进行对比. 为保证对比的公平性, 所有基线模型均在统一环境和设置上进行复现与训练. 具体采用的基线情况如下:

● 传统深度学习方法. 此类方法主要依赖于卷积神经网络或 Transformer 架构提取面部视觉特征及多模态特征. 对比模型包括 Video InceptionNet[32]、C3D + LSTM[33]、ResNet + LSTM[34]、Swin-Transformer[35]、ViViT[36]、EfficientNet + LSTM[37]、EfficientNet + Bi-LSTM[37]、MGAFR[19]、VisioPhysioENet[20]、MIST[21] 以及 FANN[38]. 在数据输入层面, 采用分段采样策略, 即每个视频样本被划分为 16 个片段, 每段抽取 1 帧, 以消减时序冗余.

● 视觉语言模型. 为探究大规模预训练模型在特定领域的迁移能力, 本文选取近几年发布的有代表性的视觉语言模型, 包括 Qwen3-VL[28]、Video-LLaVA[39] 及 LLaVA-NeXT[40], 并将其应用在学生参与度评估任务中.

针对 VisioPhysioENet 等需要生理信号的方法, 由于 DAISEE 和 EmotiW2023 数据集未提供原生的血容量脉搏 (BVP) 等接触式生理信号, 本文严格遵照 VisioPhysioENet 原论文中的非接触式方案, 采用远程光电容积描记法 (remote photoplethysmography, rPPG)(具体为 CHROM (chrominance-based) 方法) 从人脸视频中提取远程心率与脉搏信号作为生理表征输入. 这保证了对比方法的客观性与原作者思路的公平复现. 此外, DIPSER 数据集的低采样帧率导致 rPPG 无法提取出稳定周期的生理频带, 因此该方法在该数据集上的表现未能在表 2 中提供 (以“—”标注). 值得补充的是, 在对比模型的选取上, 本文亦纳入了针对多模态情感识别 (multimodal emotion recognition, MER) 任务的近期前沿模型 (如 MGAFR 与 MIST). 从理论层面看, SEP 与 MER 在面部动作编码系统 (facial action coding system, FACS) 的基础表征层面高度重合; 从技术层面看, 多传感器异构融合与时序特征插补是两者的共性挑战. 因此, 在统一的实验基准下衡量前沿 MER 架构, 足以作

为检验本文跨模态特征融合效能的有力参照.

4.4 对比实验与结果分析

为全面评估所提模型 VLM-SEP 的有效性, 本文在 DAISEE、EmotiW2023 及 DIPSER 三个学生参与度公开数据集上进行广泛的定量对比实验. 实验结果以多次独立运行的均值 ± 标准差形式报告, 加粗表示整体最佳性能, 下划线表示基线方法中的最佳性能, "—"表示该方法在对应数据集上无法评估, 详细数据如表 2 所示. 整体结果表明, VLM-SEP 在多个评估维度上均展现出显著的性能优势. 首先, VLM-SEP 的整体预测性能优秀, 在 DAISEE、EmotiW2023 及 DIPSER 数据集上分别取得 59.01% ± 0.20%、61.08% ± 0.10% 以及 65.82% ± 0.20% 的准确率, 在所有参评模型中均达到最优水平. 值得注意的是, 在样本分布极度不均衡且环境噪声复杂的 DAISEE 数据集上, VLM-SEP 仍能保持领先, 证明模型在非受控环境下的鲁棒性.

此外, 相比于仅依赖时空特征提取的传统深度网络 (如 Video InceptionNet、C3D + LSTM 等), VLM-SEP 展现出显著优势, 在 DIPSER 上, 较 EfficientNet + Bi-LSTM 提升 1.28%. 这反映出传统模型在面对"语义断层"问题时, 仅凭像素级的特征拟合难以捕获参与度背后的逻辑关联, 而引入多模态深层语义特征进行融合, 比单纯拟合时空视觉模式更能深刻揭示学生行为特征的内在逻辑. 在 DAISEE 和 EmotiW2023 上, VLM-SEP 较最佳基线 MGAFR 分别提升 0.65% 和 1.69%, 这归功于本文提出的知识引导推理模块, 它不仅引入了文本, 更通过教育学先验规则对跨模态融合进行显式约束, 避免多模态信息的无效堆叠. 最后, 实验结果揭示一个重要现象, 即直接将开源的通用视觉语言模型 (如 Video-LLaVA (7B)、LLaVA-NeXT (7B) 和 Qwen3-VL (8B)) 应用于细粒度的学生参与度预测时, 性能出现严重衰减, 甚至远不及传统深度学习基线. 分析其根源, 通用大模型虽然具备广博的知识, 但其注意力机制在处理长视频序列时存在感知钝化现象, 难以精准锁定视线偏移等瞬时微表情. 而 VLM-SEP 通过注入教育学先验知识, 执行多层次特征解耦与文本化推理, 有效解决了通用大模型难以直接捕捉特定领域微观特征的问题, 实现从通用视觉问答到精准行为判别的重要跨越.

4.5 消融实验

为验证所提 VLM-SEP 框架中各核心组件的有效性与必要性, 本文设计并开展全面的消融实验. 通过逐步移除或替换关键模块并量化其对整体预测精度的影响, 旨在深入解析各组件的具体贡献. 为确保评估的公平性与可靠性, 所有变体实验均在严格统一的数据划分、超参数配置及硬件环境下进行.

实验设计主要围绕两个核心维度展开: 一是模态必要性分析, 通过构建"仅视觉"与"仅文本"基线模型, 独立评估单一模态特征对预测性能的贡献; 二是特征融合机制有效性验证, 构建"简单融合"(直接拼接或逐元素相加)、"无知识引导模块"、"单向视觉引导文本"及"单向文本引导视觉"四个对照组. 此设计旨在探究双向交叉注意力机制在跨模态信息交互中的优越性. 具体实验结果详见表 3, 关键结论分析如下:

- 多模态特征融合的必要性. 实验结果表明, 多模态协同方案的综合性能显著优于单一模态分支. 以 DAISEE 数据集为例, 完整模型的准确率达到 59.01%, 较"仅视觉"和"仅文本"基线分别提升 1.05 和 5.97 个百分点, 且平均绝对误差降至最优的 0.418 6. 这一结果表明, 视觉生理特征与解耦后的文本语义信息之间存在显著的特征互补效应.
- 双向交叉注意力机制的优越性. 相较于简单拼接、单向交互等, 本文提出的双向交叉注意力机制能够更高效地实现跨模态特征的深度对齐. 在 DAISEE 和 EmotiW2023 数据集上, 均取得最优的 F1 分数 (分别为 0.555 4 和 0.597 0), 证明了双向交互在捕捉复杂多模态映射关系中的关键作用.
- 知识引导模块的必要性. 通过"无知识引导模块"消融实验 (即将第 3.2 节的知识引导推理替换为不含教育心理学规则约束的通用文本描述), 可以明确量化该模块的边际贡献. 在 DAISEE 数据集上, 移除知识引导后准确率从 59.01% 下降至 56.79%, F1 分数从 0.555 4 降至 0.525 9. 在 EmotiW2023 上, 准确率从 61.08% 降至 60.04%. 在 DIPSER 上, 准确率从 65.82% 降至 65.15%. 这些结果充分证明教育心理学先验知识作为归纳偏置的有效性与必要性: 知识引导模块通过注入领域规则约束, 使模型在跨模态推理过程中具备更强的语义判别能力.
- 模型的综合鲁棒性与泛化能力. 尽管对照组在某些局部指标上表现出微弱优势 (例如, EmotiW2023 中"单向视觉引导文本"的准确率略高, 或 DIPSER 数据集中"仅文本"的平均绝对误差达到 0.365 6), 但纵观三大数据集的各项核心指标综合表现, 完整模型展现出最佳的鲁棒性. 该机制有效缓解单向交互或简单融合在特定数据分布下容易出现的性能显著衰退问题, 具备更强的跨领域泛化潜力.

表 2　在三个学生数据集上 VLM-SEP 与传统方法的准确率对比 (%)
Table 2　Accuracy comparison between VLM-SEP and traditional methods on three student datasets (%)

模型	输入	DAISEE	EmotiW2023	DIPSER
Video InceptionNet	视频帧	53.98 ± 0.30	52.05 ± 0.30	63.63 ± 0.20
C3D + LSTM	视频帧	57.08 ± 0.20	51.44 ± 0.20	63.44 ± 0.25
ResNet + LSTM	视频帧	57.62 ± 0.15	54.05 ± 0.15	64.11 ± 0.30
Swin-Transformer	视频帧	55.62 ± 0.30	52.52 ± 0.30	63.44 ± 0.20
ViViT	视频帧	57.03 ± 0.25	54.17 ± 0.25	63.99 ± 0.30
EfficientNet + LSTM	视频帧	56.98 ± 0.25	58.13 ± 0.25	63.56 ± 0.20
EfficientNet + Bi-LSTM	视频帧	57.33 ± 0.25	58.87 ± 0.35	64.54 ± 0.20
FANN	视频帧	58.08 ± 0.30	57.91 ± 0.40	64.17 ± 0.20
VisioPhysioENet	视频帧 + 生理信号	56.64 ± 0.10	51.30 ± 0.20	—
MGAFR	视频帧 + 文本	58.36 ± 0.10	59.39 ± 0.10	—
MIST	视频帧 + 文本	57.22 ± 0.20	57.86 ± 0.20	—
Video-LLaVA (7B)	视频帧 + 文本	49.97 ± 0.00	29.73 ± 0.00	48.24 ± 0.00
LLaVA-NeXT (7B)	视频帧 + 文本	45.17 ± 0.00	24.67 ± 0.00	21.11 ± 0.00
Qwen3-VL (8B)	视频帧 + 文本	50.24 ± 0.00	51.30 ± 0.00	55.23 ± 0.00
VLM-SEP	视频帧 + 文本	**59.01 ± 0.20**	**61.08 ± 0.10**	**65.82 ± 0.20**

● 跨数据集泛化性. 在主对比实验 (表 2) 中, VLM-SEP 在 DAISEE、EmotiW2023 和 DIPSER 三个数据集上的准确率均达到所有参评方法中的最优水平. 在消融实验 (表 3) 中, 个别变体在特定数据集的局部指标上出现微弱优势 (如“单向视觉引导文本”在 EmotiW2023 上准确率为 61.92%, 略高于完整模型的 61.08%), 但其在其他数据集上的性能出现明显衰退 (DAISEE 上仅为 58.08%). 这种现象恰好说明双向交叉注意力机制在不同数据分布下的稳健性: 完整模型牺牲在单一数据集上的微弱局部优势, 换取跨三个数据集的整体最优综合性能, 体现更强的泛化能力.

4.6　特征可视化分析

为了直观验证所提多层次特征解耦模块的有效性, 本实验采用 t-SNE 算法对解耦前后的特征分布进行可视化对比分析. 观察图 4 中三组数据集的实验结果可以发现, 解耦前的原始图像特征在空间中呈现出显著的“碎片化”分布, 样本形成大量孤立的局部小簇. 通过深入分析发现, 这些小簇主要由视频的背景、光照及个体身份信息主导, 导致不同参与度等级的样本在全局范围内高度纠缠, 缺乏清晰的类间边界. 这表明原始视觉表征包含大量与参与度任务无关的冗余信息.

经过特征解耦后的语义特征展现出截然不同的分布形态, 原本碎片化的结构消失, 特征空间实现从感知物理特征向理解行为语义的转变. 在三个数据集上, 不同参与度等级的样本均表现出明显的全局演变趋势与语义聚类效应: 代表低参与度的紫色/蓝色区域与代表高参与度的黄色区域在空间上实现有效分离, 且呈现出连续的序数逻辑分布. 特别是对于高参与度样本, 在解耦后形成更为紧凑且独立的簇群, 显著降低后续推理模块的分类难度.

综合三组数据集的可视化表现, 本文提出的特

表 3　消融实验
Table 3　Ablation study

消融模块	DAISEE			EmotiW2023			DIPSER		
	准确率 (%)	F1 分数	平均绝对误差	准确率 (%)	F1 分数	平均绝对误差	准确率 (%)	F1 分数	平均绝对误差
仅视觉	57.96	0.533 0	0.430 9	59.78	0.585 7	0.563 3	64.90	0.611 6	0.402 7
仅文本	53.04	0.515 9	0.485 4	60.43	0.537 7	0.533 5	63.44	0.492 5	**0.365 6**
简单融合	58.72	0.546 0	0.422 7	59.78	0.566 8	0.544 7	65.70	**0.635 0**	0.376 5
无知识引导模块	56.79	0.525 9	0.441 5	60.04	0.580 1	0.499 1	65.15	0.550 3	0.357 1
单向视觉引导文本	58.08	0.547 2	0.430 9	**61.92**	0.596 9	0.496 3	64.78	0.531 9	0.366 8
单向文本引导视觉	57.73	0.479 2	0.445 6	60.15	0.584 4	0.522 3	65.39	0.626 1	0.389 9
完整模型 (VLM-SEP)	**59.01**	**0.555 4**	**0.418 6**	61.08	**0.597 0**	0.533 5	**65.82**	0.610 8	0.374 7

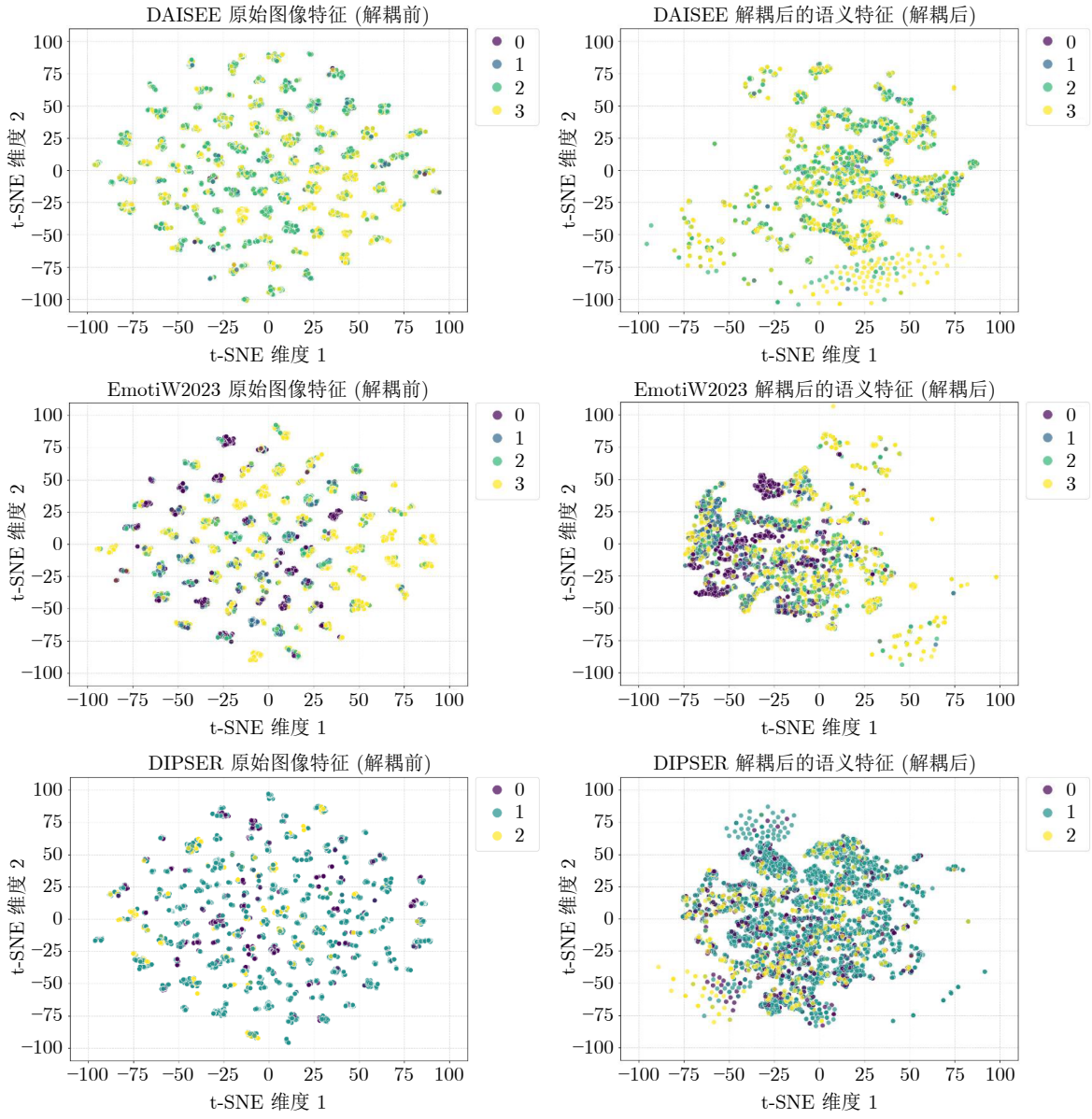


图 4 特征解耦前后样本分布对比

Fig. 4 Comparison of sample distributions before and after feature decoupling

征解耦方法成功抑制非任务相关的噪声干扰, 打破特征对特定个体或环境的过度依赖. 解耦后的表征能够更精准地刻画与学习参与度相关的核心行为语义, 证明该模块在跨场景、跨个体任务中具有极强的鲁棒性与特征提取的纯净度, 为实现高精度的学生参与度分析奠定坚实的表征基础.

4.7 视觉–文本相关性分析

为了深层探究视觉语言模型提取的结构化文本特征与原始视觉帧特征之间的语义相关性, 本文引入双向交叉注意力可视化验证实验, 旨在验证模型是否能将抽象的教育心理学文本描述 (如“眼睑下垂”、“视线偏离”) 映射到对应的视频时序帧上.

可视化矩阵中的数值源于 Transformer 架构中的缩放点积注意力机制. 令视觉特征序列为 $\boldsymbol{f}_v \in \mathbf{R}^{T\times D}$, 文本特征序列为 $\boldsymbol{f}_t \in \mathbf{R}^{T\times D}$ (其中 $T=16$ 为序列帧数, $D=2\ 048$ 为特征维度). 在文本查询视觉的过程中, 文本特征投影为查询矩阵 Q, 将视觉特征投影为键矩阵 K:

$$Q = \boldsymbol{f}_t \boldsymbol{W}_Q^t, \quad K = \boldsymbol{f}_v \boldsymbol{W}_K^v \tag{18}$$

热力图中第 i 行、第 j 列的数值 $A_{i,j}$ 代表第 i 条文本描述与第 j 个视频帧之间的相关性得分, 其计算公式如下:

$$A_{i,\,j} = \mathrm{Softmax}\left(\frac{Q_i K_j^{\mathrm{T}}}{\sqrt{D_p}}\right) = \frac{\exp(\frac{Q_i K_j^{\mathrm{T}}}{\sqrt{D_p}})}{\sum_{l=1}^{T}\exp(\frac{Q_i K_l^{\mathrm{T}}}{\sqrt{D_p}})} \tag{19}$$

其中, $\sqrt{D_p}$ 为缩放因子, 用于防止点积数值过大导致梯度消失. 通过 Softmax 函数的归一化, 每一行的权重总和为 1, 直观地展示模型在处理特定文本语义时对视觉序列的聚焦分布.

图 5 展示在“无参与”、“极低参与”、“中等参与”及“高度参与”四种典型状态下的注意力分布. 从图中可以看出以下两点现象:

- 对角线一致性. 在所有等级的热力图中, 矩阵均呈现出极为显著的对角线高亮特征 (即当 $i = j$ 时 $A_{i,\,j}$ 取得最大值). 这证明了模型实现精准的视觉–文本语义对齐, 即第 i 帧的视觉行为特征被模型准确地解耦并映射到第 i 条文本语义中, 未出现时序错位.
- 语义稳健性. 值得注意的是, 即使在“无参与”状态下, 模型依然能保持清晰的对角线分布. 这说明模型捕捉到的相关性并不依赖于学生的表现优劣, 而是源于对行为特征的通用理解能力.

综上所述, 交叉注意力的可视化结果证实了本文提出的 VLM-SEP 模型具备较好的跨模态语义映射能力, 能够为后续的参与度等级判定提供高度可信的时空特征支撑, 同时也增强深度学习模型在教育场景应用中的可解释性.

4.8 混淆矩阵与边界分析

为了直观展示模型对边界样本与专家分歧样本的处理能力, 图 6 给出了 VLM-SEP 方法在三个数据集上的混淆矩阵. 在基于多专家标注且存在不可避免主观分歧的 DIPSER 数据集上, 混淆矩阵呈现出极具教育学启发性的分布规律——模型的误分类主要集中在相邻类别之间 (如将真实的“中等参与”判定为“高度参与”), 而跨象限的极端误判 (如“极低参与”被误判为“高度参与”) 所占比例微乎其微. 这深度契合参与度这一标量属性本质上是一种连续过渡的心理状态. 此外, 针对边界样本, 由于不同教育专家对于“学生微弱的视线下移”究竟属于高度参与还是极低参与存在阈值分歧, VLM-SEP 通过引入基于通用语料的先验规则空间与跨模态文本语义的融合, 能够自适应平滑个别专家的极端标定偏差, 进而稳定收敛至更接近综合预期的参与度判定. 这证明了模型除了学习统计关联, 还具备强大的抗标签噪声鲁棒性.

4.9 案例分析

为了更加清晰地观察 VLM-SEP 模型是如何对文本特征进行提取分析以及与最终分类的关联性和可解释性, 本实验通过抽取具有代表性的案例展示模型预测的全过程. 如图 7 列出的两个具体案例所示, VLM-SEP 模型首先进行多层次特征解耦, 将原始视频流精细化地转化为结构化的特征描述序列, 包括视线方向、眼睑状态、头部姿态及嘴角动作等. 例如, 在“高度参与”案例中, 模型捕捉到“视线居中、眼睑正常、头部直立”的稳定特征流; 在“极低参与”案例中, 则准确识别出“眼睑下垂、视线偏移”等低唤醒特征. 这一跨模态转换过程将抽象的视觉信号具象化, 为后续的逻辑推理提供可靠的感知基础. 在获取结构化特征后, 模型通过知识引导的注意力推理模块模拟教育专家的判别逻辑, 对学生在特定视频帧的参与状态进行深度推理. 例如, 针对极低参与度学生, 推理模块通过“低唤醒特征 → 情感淡漠状态 → 生理上的眼睑/目光下垂 → 疲劳或注意力分散”的逻辑链条, 逐步推导出“极低参与”的最终结论. 这种基于教育心理学逻辑的引导式推理, 显著提升了模型在实际教学评估中的可信度与应用价值.

此外, 通过与传统端到端深度学习模型进行可视化定性分析对比, 本文进一步验证了 VLM-SEP 在复杂教学场景下的判别优势. 实验中抽取图 8 所示的一组案例——学生 A 在视频中表现出视线下垂及嘴角下压的特征. 若采用传统深度学习模型 (如 ResNet + LSTM 等), 其注意力机制受限于纯数据驱动, 仅根据热力图反馈将注意力集中于面部额头位置, 由于忽视了“嘴角下压”这一反映学生状态不佳的关键负面特征, 导致产生浅层偏见, 最终给出错误的“高度参与”判定, 如图 8 所示. 相比之下, VLM-SEP 展现更精准的推理逻辑: 首先, 系统通过多层次参与度特征解耦, 精准捕捉到学生视线向下、眼睑下垂、嘴角下压等细微动作; 随后, 在“知识引导的注意力推理”阶段, 模型结合教育心理学先验知识进行逻辑演绎, 识别出该学生虽然保持较稳定的行为参与, 但同时存在少量视线方向偏移和嘴角下垂等低参与度特征. 最终, 通过跨模态融合决策框架, 系统纠正传统视觉模型的误判, 得出与实际情况相符的“中度参与”判定.

5 结束语

本文针对通用视觉语言模型在学生参与度预测

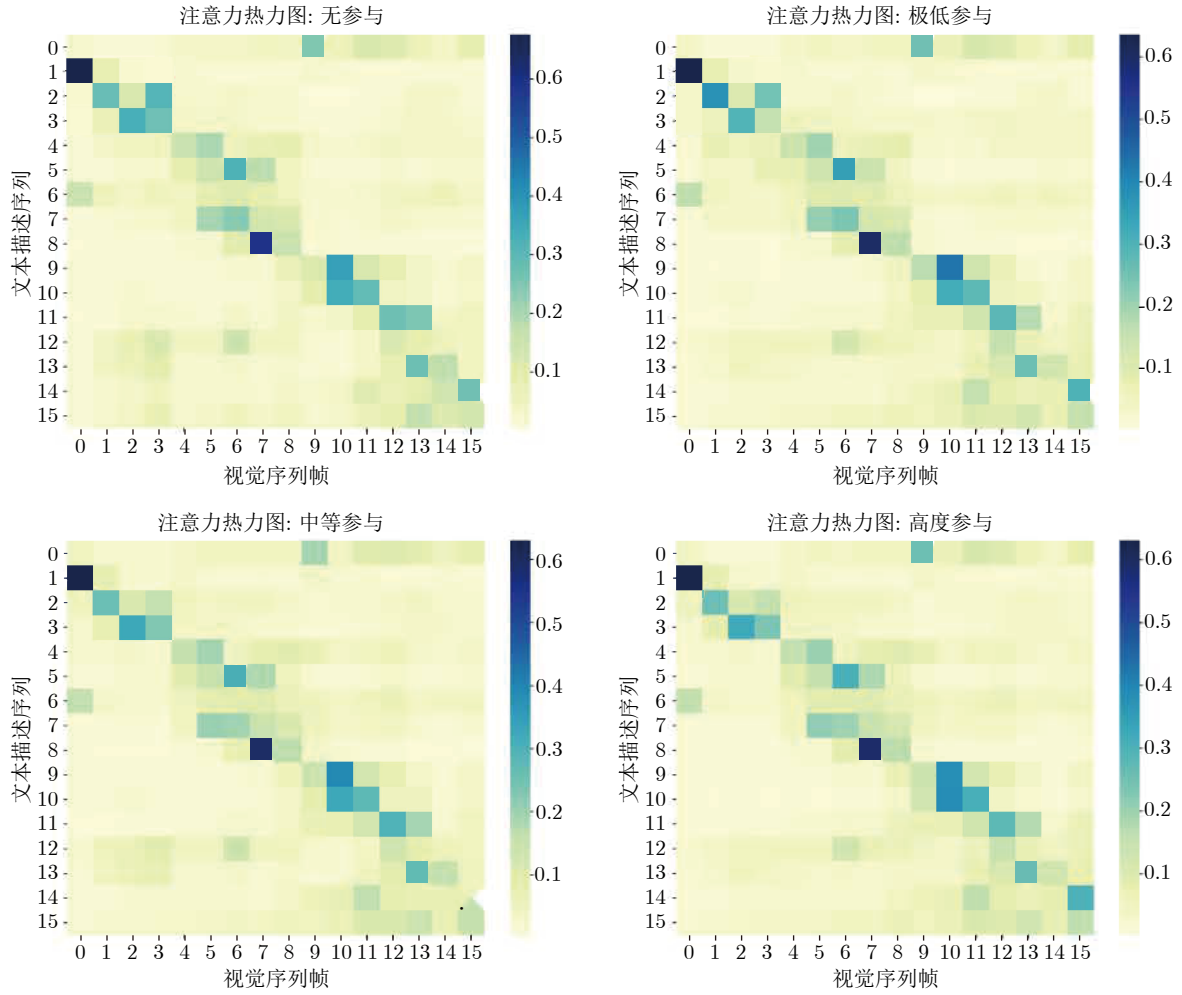


图 5 文本特征与视频特征的时序相关性热力图

Fig. 5 Temporal correlation heatmap between text and video features

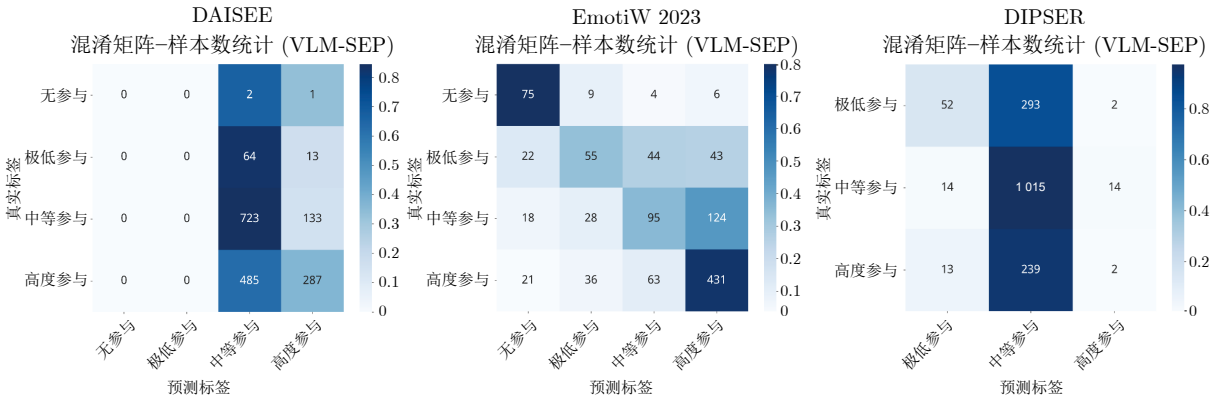


图 6 VLM-SEP 方法在三个数据集上的混淆矩阵

Fig. 6 Confusion matrices of the VLM-SEP method on the three datasets

任务中直接迁移时所面临的语义鸿沟与微观特征感知瓶颈, 提出一种基于视觉语言模型的多模态学生参与度预测方法. 该方法构建一种递进式的多模态推理架构: 首先, 通过多层次参与度特征解耦, 从面部表情与动作姿态中提取结构化视觉特征; 其次, 引入教育心理学先验知识作为推理规则, 建立离散视觉特征与参与度状态自然语言描述间的显式映射; 最后, 通过双向交叉注意力机制进行跨模态融合决

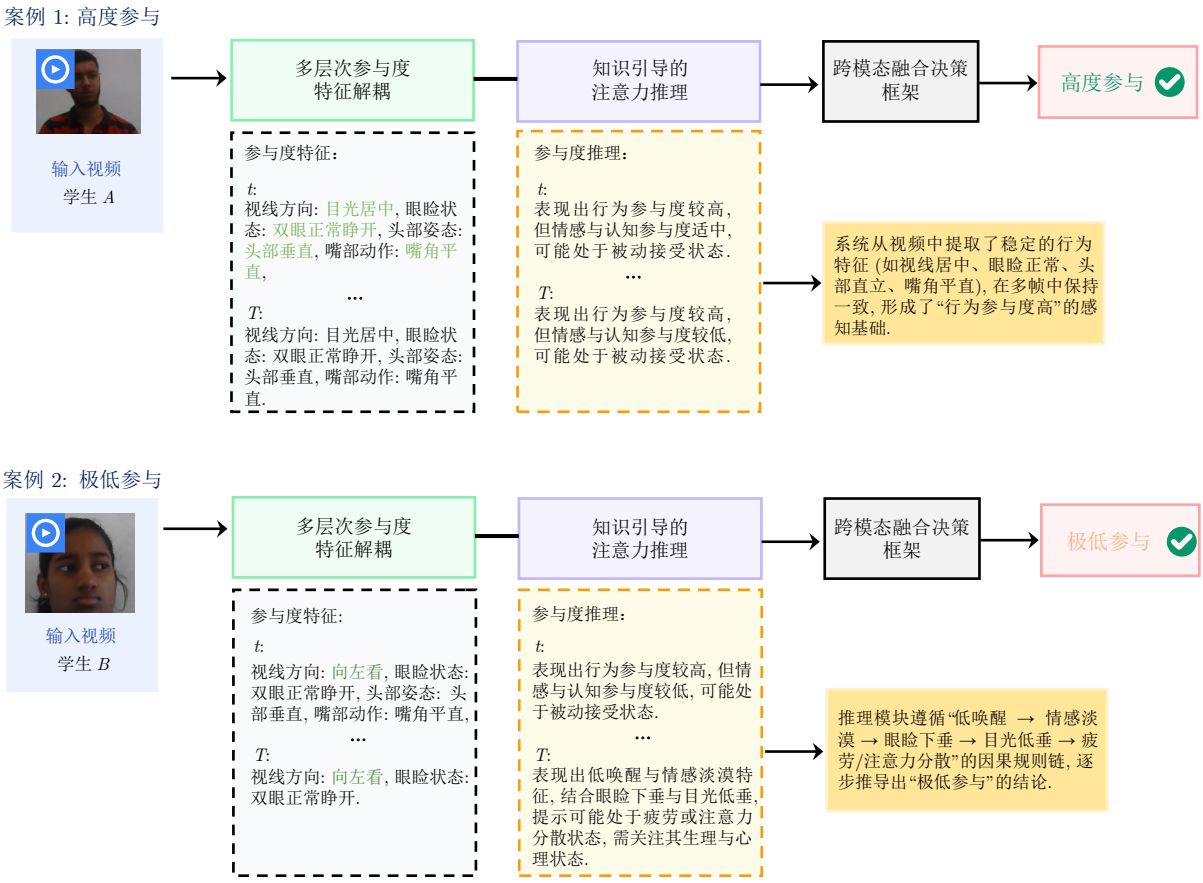


图 7　案例分析

Fig. 7　Case analysis

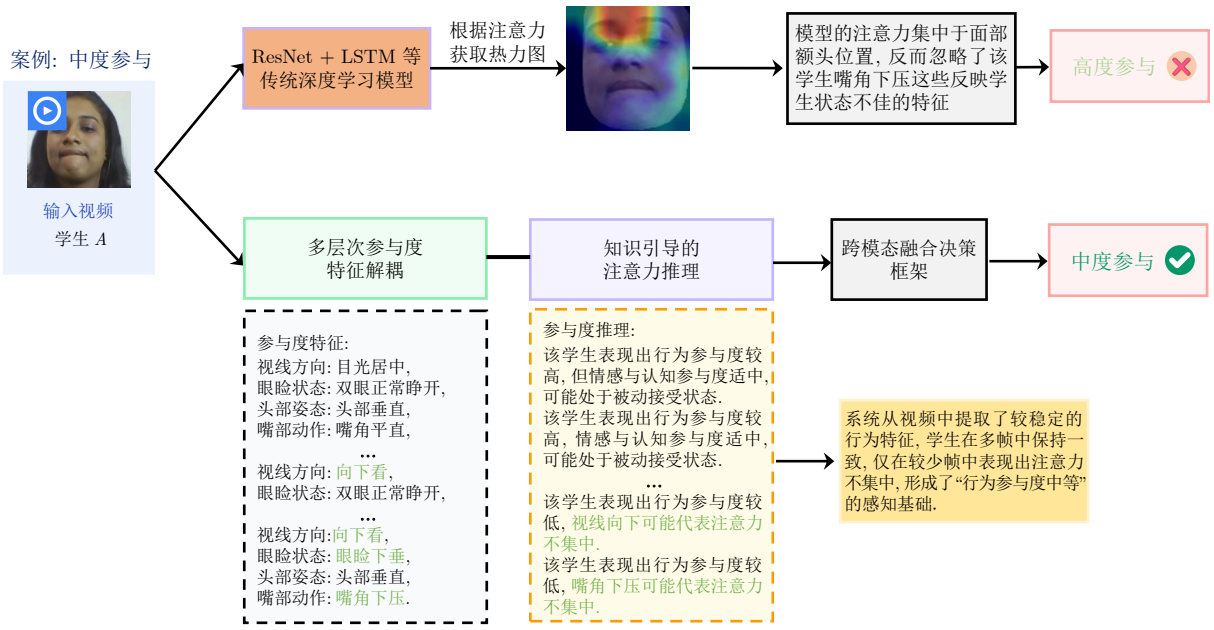


图 8　定性分析

Fig. 8　Qualitative analysis

策, 实现视觉直观表征与文本逻辑推理的深度耦合, 从而提升判别精度. 在 DAISEE、EmotiW2023 及 DIPSER 三个公开数据集上的广泛实验表明, VLM-SEP 显著优于现有的传统深度学习模型及通用视觉语言大模型. 此外, 消融实验与可视化分析进一步验证了多模态双向交互机制在捕捉复杂数据映射关系、增强模型鲁棒性以及提升预测结果可解释性方面的核心优势.

尽管 VLM-SEP 在自动化参与度评估中展现出优异的性能与可解释性, 该领域仍具进一步探索的潜力. 首先, 现有框架主要局限于视觉视频流与文本逻辑推理的融合, 未来研究可引入音频交互信号或生理监测数据 (如心率、血容量脉搏等), 以构建多维度的学习者状态评估体系; 其次, 针对通用视觉语言模型参数量大、计算成本高的问题, 未来可结合知识蒸馏或模型量化技术, 开发适用于非受控在线教育场景及边缘计算设备的轻量级实时推理系统; 最后, 考虑到学生在参与度表达与微表情特征上存在显著个体差异, 未来工作将探索长周期时序动态建模方法, 并为不同学习者引入个性化行为基线标定机制, 以更精准地追踪学习周期内的认知与行为演变规律.

参考文献

1 Xiao Jian-Li, Huang Xing-Yu, Jiang Fei. A review of large language models in intelligent education. *CAAI Transactions on Intelligent Systems*, 2025, **20**(5): 1054−1070
(肖建力, 黄星宇, 姜飞. 智慧教育中的大语言模型综述. 智能系统学报, 2025, **20**(5): 1054−1070)

2 Doherty K, Doherty G. Engagement in HCI: Conception, theory and measurement. *ACM Computing Surveys*, 2018, **51**(5): 1−39

3 D′Mello S K. *Improving Student Engagement in and With Digital Learning Technologies*. Cham: Springer, 2021. 79–104

4 Wei Y T, Wang J X, Yang H H, Shi Y H, Zhou G W, Li X. Research on the influence of students′ engagement in blended synchronous learning. In: Proceedings of the 2023 International Symposium on Educational Technology. Hong Kong, China: IEEE, 2023. 37–41

5 Fredricks J A, Blumenfeld P C, Paris A H. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, 2004, **74**(1): 59−109

6 Fredricks J A, Filsecker M, Lawson M A. Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, 2016, **43**: 1−4

7 Li T, Zhu A. How online classroom interaction tool change the teaching and learning mode under the traditional computer-assisted instruction? In: Proceedings of the International Conference on Computer Science and Educational Informatization. Kunming, China: IEEE, 2019. 25–29

8 Smith J, Schreder K. Are they paying attention, or are they shoe-shopping? Evidence from online learning. *International Journal of Multidisciplinary Perspectives in Higher Education*, 2020, **5**(1): 200−209

9 Mo Yuan-Jiao. Classroom student engagement prediction method based on cross-branch attention learning. *Industrial Control Computer*, 2025, **38**(1): 118−120
(莫元娇. 基于跨分支注意力学习的课堂学生参与度预测方法. 工业控制计算机, 2025, **38**(1): 118−120)

10 He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016. 770–778

11 Geng L, Xu M, Wei Z Q, Zhou X Z. Learning deep spatiotemporal feature for engagement recognition of online courses. In: Proceedings of the Symposium Series on Computational Intelligence. Xiamen, China: IEEE, 2019. 442–447

12 Wo Yan, Liang Ji-Yun, Han Guo-Qiang. Cross-modal face retrieval method based on metric learning. *Journal of South China University of Technology* (*Natural Science Edition*), 2022, **50**(6): 1−9
(沃焱, 梁籍云, 韩国强. 基于度量学习的跨模态人脸检索方法. 华南理工大学学报 (自然科学版), 2022, **50**(6): 1−9)

13 Zhang H, Xiao X F, Huang T, Liu S Y, Xia Y, Li J. A novel end-to-end network for automatic student engagement recognition. In: Proceedings of the 9th International Conference on Electronics Information and Emergency Communication. Beijing, China: IEEE, 2019. 342–345

14 Liao J C, Liang Y, Pan J H. Deep facial spatiotemporal network for engagement prediction in online learning. *Applied Intelligence*, 2021, **51**(10): 6609−6621

15 Liu Z, Kong W Z, Peng X, Yang Z K, Liu S, Liu S Q, et al. Dual-feature-embeddings-based semisupervised learning for cognitive engagement classification in online course discussions. *Knowledge-Based Systems*, 2023, **259**: Article No. 110053

16 Huang T, Mei Y S, Zhang H, Liu S Y, Yang H L. Fine-grained engagement recognition in online learning environment. In: Proceedings of the 9th International Conference on Electronics Information and Emergency Communication. Beijing, China: IEEE, 2019. 338–341

17 Singh M, Hoque X, Zeng D H, Wang Y N, Ikeda K, Dhall A. Do I have your attention: A large scale engagement prediction dataset and baselines. In: Proceedings of the 25th ACM International Conference on Multimodal Interaction. New York, USA: ACM, 2023. 174–182

18 Vedernikov A, Kumar P, Chen H Y, Seppänen T, Li X B. TCCT-Net: Two-stream network architecture for fast and efficient engagement estimation via behavioral feature signals. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE, 2024. 4723–4732

19 Deng Y Y, Bian J T, Wu S S, Lai J H, Xie X H. Multiplex graph aggregation and feature refinement for unsupervised incomplete multimodal emotion recognition. *Information Fusion*, 2025, **114**: Article No. 102711

20 Singh A, Verma N, Goyal K, Singh A, Kumar P, Li X B. VisioPhysioENet: Multimodal engagement detection using visual and physiological signals. arXiv preprint arXiv: 2409.16126, 2024.

21 Boitel E, Mohasseb A, Haig E. MIST: Multimodal emotion recognition using DeBERTa for text, Semi-CNN for speech, ResNet-50 for facial, and 3D-CNN for motion analysis. *Expert Systems With Applications*, 2025, **270**: Article No. 126236

22 Liu Y H, Kuang Z Y, Zhang H Y, Li C, Li F F, Ding X H. PRADA: Prompt-guided representation alignment and dynamic adaption for time series forecasting. *Knowledge-Based Systems*, 2025, **318**: Article No. 113478

23 Zou X, Li X H, Hu P, Dong M. MrBalance: A framework for enhancing event causality identification in multi-agent debates via role assignment. *Knowledge-Based Systems*, 2025: Article No. 114470

24 Qi X Y, Zeng Y, Xie T H, Chen P Y, Jia R X, Mittal P, et al. Fine-tuning aligned language models compromises safety, even when users do not intend to! arXiv preprint arXiv: 2310.03693, 2023.

25 Heilporn G, Raynault A, Frenette É. Student engagement in a higher education course: A multidimensional scale for different course modalities. *Social Sciences & Humanities Open*, 2024, **9**: Article No. 100794

26 Fredricks J A, McColskey W. The measurement of student en-

gagement: A comparative analysis of various methods and student self-report instruments. *Handbook of Research on Student Engagement*. Boston, MA: Springer US, 2012. 763–782

27 Blikstein P. Multimodal learning analytics. In: Proceedings of the 3rd International Conference on Learning Analytics and Knowledge. New York, USA: ACM, 2013. 102–106

28 Yang A, Li A F, Yang B S, Zhang B C, Hui B Y, Zheng B, et al. Qwen3 technical report. arXiv preprint arXiv: 2505.09388, 2025.

29 Gupta A, D′Cunha A, Awasthi K, Balasubramanian V. DAiSEE: Towards user engagement recognition in the wild. arXiv preprint arXiv: 1609.01885, 2016.

30 Dhall A, Singh M, Goecke R, Gedeon T, Zeng D H, Wang Y N, et al. EmotiW2023: Emotion recognition in the wild challenge. In: Proceedings of the 25th ACM International Conference on Multimodal Interaction. New York, USA: ACM, 2023. 746–749

31 Marquez-Carpintero L, Suescun-Ferrandiz S, Álvarez C L, Fernandez-Herrero J, Viejo D, Roig-Vila R, et al. DIPSER: A dataset for in-person student engagement recognition in the wild. arXiv preprint arXiv: 2502.20209, 2025.

32 Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco, CA, USA: AAAI, 2017. 4278–4284

33 Parmar P, Morris B. Action quality assessment across multiple actions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa Village, HI, USA: IEEE, 2019. 1468–1476

34 Abedi A, Khan S S. Improving state-of-the-art in detecting student engagement with ResNet and TCN hybrid network. In: Proceedings of the 18th Conference on Robots and Vision. Burnaby, British Columbia, Canada: IEEE, 2021. 151−157

35 Liu Z, Lin Y T, Cao Y, Hu H, Wei Y X, Zhang Z, et al. Swin Transformer: Hierarchical vision Transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021. 10012–10022

36 Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A video vision Transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, QC, Canada: IEEE, 2021. 6836–6846

37 Selim T, Elkabani I, Abdou M A. Students engagement level detection in online e-learning using hybrid EfficientNetB7 together with TCN, LSTM, and Bi-LSTM. *IEEE Access*, 2022, **10**: 99573−99583

38 Wang H, Sun H M, Zhang W L, Chen Y X, Jia R S. FANN: A novel frame attention neural network for student engagement recognition in facial video. *The Visual Computer*, 2025, **41**: 6011−6025

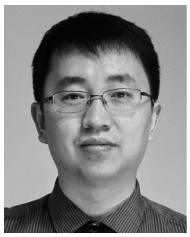
39 Lin B, Ye Y, Zhu B, Cui J X, Ning M N, Jin P, et al. Video-LLaVA: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Miami, Florida, USA: ACL, 2024. 5971–5984

40 Li B, Zhang K C, Zhang H, Guo D, Zhang R R, Li F, et al. LLaVA-NeXT: Stronger LLMs supercharge multimodal capabilities in the wild. arXiv preprint arXiv: 2408.01073, 2024.

沈双宏 合肥综合性国家科学中心人工智能研究院副研究员. 主要研究方向为数据挖掘, 智能教育.
E-mail: shshen@iai.ustc.edu.cn
(**SHEN Shuang-Hong** Associate researcher at the Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. His research interests include data mining and intelligent education.)

秦子轩 安徽大学硕士研究生. 主要研究方向为计算机视觉, 多模态融合及视觉语言模型应用.
E-mail: wa24201053@stu.ahu.edu.cn
(**QIN Zi-Xuan** Master student at Anhui University. His research interests include computer vision, multimodal fusion, and the application of vision-language models.)

苏 喻 合肥师范学院教授. 主要研究方向为自然语言理解, 智慧教育.
E-mail: yusu@hfnu.edu.cn
(**SU Yu** Professor at Hefei Normal University. His research interests include natural language understanding and intelligent education.)



王士进 科大讯飞股份有限公司正高级工程师, 副总裁. 主要研究方向为认知智能及智慧教育.
E-mail: sjwang3@iflytek.com
(**WANG Shi-Jin** Professor-level senior engineer, vice president at iFLYTEK Co., Ltd. His research interests include cognitive intelligence and smart education.)

黄振亚 中国科学技术大学计算机科学与技术学院副教授. 主要研究方向为数据挖掘及认知推理.
E-mail: huangzy@ustc.edu.cn
(**HUANG Zhen-Ya** Associate professor at the School of Computer Science and Technology, University of Science and Technology of China. His research interests include data mining and cognitive reasoning.)

孙登第 安徽大学人工智能学院教授. 主要研究方向为计算机视觉, 机器学习与深度学习. 本文通信作者.
E-mail: sundengdi@ahu.edu.cn
(**SUN Deng-Di** Professor at the School of Artificial Intelligence, Anhui University. His research interests include computer vision, machine learning, and deep learning. Corresponding author of this paper.)