Advances in Education 教育进展, 2026, 16(4), 1224-1234
Published Online April 2026 in Hans. https://www.hanspub.org/journal/ae
https://doi.org/10.12677/ae.2026.164772

Hans汉斯

基于多模态融合驱动的智慧课堂感知系统实验方法设计

李禹衡，丁晓铭*，张　重，韩　亮

天津师范大学人工智能学院，天津

收稿日期：2026年3月17日；录用日期：2026年4月15日；发布日期：2026年4月27日

摘　要

针对人工智能师范专业实验教学中单一模态训练局限、技术与教育场景脱节，以及学生课堂整体感知能力培养不足的问题，本文设计并实现了基于语音－视觉双模态融合的智慧课堂感知系统及对应实验。实验以“技术融合－场景适配－能力验证”为核心，整合Whisper、YOLO26、改进型ViT及大语言模型，构建全流程实验体系，实现教学阶段划分、学生个体识别与追踪、动作识别及多模态数据融合，输出课堂感知综合报告。结合前文实验流程与验证，结果表明，该方案规范可行，突破传统实验局限，可帮助师范生巩固核心技术、提升工程实践能力，培养课堂感知与场景适配能力，为智能教育复合型师资培养提供实践载体，也为相关专业实验教学体系优化提供参考。

关键词

多模态融合，智慧课堂，人工智能师范，实验教学，课堂感知

Design of Experimental Method for Smart Classroom Perception System Driven by Multimodal Fusion

Yuheng Li, Xiaoming Ding*, Zhong Zhang, Liang Han

School of Artificial Intelligence, Tianjin Normal University, Tianjin

Received: March 17, 2026; accepted: April 15, 2026; published: April 27, 2026

Abstract

Aiming to address the problems of single-modal training limitations, the disconnect between

*通讯作者。

文章引用: 李禹衡, 丁晓铭, 张重, 韩亮. 基于多模态融合驱动的智慧课堂感知系统实验方法设计[J]. 教育进展, 2026, 16(4): 1224-1234. DOI: 10.12677/ae.2026.164772

technology and educational scenarios, and the insufficient cultivation of students' overall classroom perception ability in the experimental teaching of AI-oriented teacher education programs, this paper designs and implements a smart classroom perception system based on audio-visual bimodal fusion, along with corresponding experiments. Centered on the framework of "technology integration-scenario adaptation-capability validation", the experiments integrate Whisper, YOLO26, an improved Vision Transformer (ViT), and a large language model to build a full-process experimental system, achieving teaching phase classification, individual student identification and tracking, action recognition, and multimodal data fusion, and outputting a comprehensive classroom perception report. In conjunction with the experimental procedures and validations described earlier, the results show that the proposed scheme is standardized and feasible, breaking through the limitations of traditional experiments. It helps pre-service teachers consolidate core technologies, enhance engineering practice skills, and cultivate classroom perception and scenario adaptation abilities. This provides a practical platform for cultivating compound talents in intelligent education and offers a reference for optimizing experimental teaching systems in related disciplines.

Keywords

Multimodal Fusion, Smart Classroom, Artificial Intelligence Normal, Experimental Teaching, Classroom Perception

1. 引言

随着人工智能技术与基础教育的深度融合，智能教育场景对具备跨学科融合能力、工程实践能力与教育认知能力的人工智能师范人才需求日益迫切。作为连接人工智能技术与教育教学实践的核心载体，智慧课堂感知系统通过整合多模态数据，实现对课堂教学状态的全面感知与精准分析，成为智能教育人才培养的重要实践载体。对于师范学院学生而言，掌握课堂整体感知能力是其未来从事智能教育相关工作的核心素养，唯有能够全面、多维度捕捉课堂教学状态，才能精准诊断教学质量、优化教学策略，真正实现技术与教育的深度融合。

当前，人工智能师范专业的实验教学多聚焦于单一技术模块的验证，如独立的图像识别、语音处理或文本分析实验，缺乏对多模态数据融合、技术与教育场景适配的系统性训练。这导致学生难以建立“技术–教育”的关联思维，无法将人工智能技术原理转化为解决教育实际问题的能力，更难以培养其课堂整体感知所需的多维度视角与综合分析能力。同时，现有实验方案对课堂教学场景的真实性覆盖不足，师生交互、课堂节奏、情感状态等核心教育要素的感知与分析维度缺失，与实际智能教育工作的需求存在差距，进一步制约了师范学生课堂整体感知能力的培养。

现有研究多聚焦单一维度，未能实现课堂场景的全方位感知，无法为师范学生提供完整示范。比如周洁等(2022)聚焦学生课堂行为检测单一维度，提出基于 OpenPose 与 CNN-10 的识别方法，虽能精准识别学生细粒度行为，但未覆盖师生交互、情绪状态等关键维度，无法还原课堂整体质量[1]。还有贾林兆等(2025)侧重课堂对话分析单一维度，提出基于 RoBERTa + BiLSTM 的分类方法，虽能实现对话精准标注，但未涉及学生非言语行为等维度，难以体现课堂深层育人价值[2]。此外，Liu 等(2025)在《Sensors》发表的相关综述研究，虽系统梳理了课堂行为识别的技术方法与研究挑战，但同样存在单一维度局限，

核心是未能将行为识别与具体教学内容相结合，仅关注行为本身，无法为师范学生提供贴合实际教学场景的完整示范[3]。

综上，无论是聚焦学生个体行为检测的相关研究，还是侧重师生言语交互分析的探索，均存在课堂感知视角单一、覆盖维度有限的共性问题，二者分别局限于学生行为分类或师生对话识别，均未实现对课堂场景多关键维度的全面覆盖，难以完整还原课堂教学真实面貌、支撑教学质量的精准诊断与优化，凸显了当前课堂感知领域研究的核心短板。在此背景下，本实验面向人工智能师范本科生，设计基于多模态融合驱动的智慧课堂感知系统实验。实验以“技术融合–场景适配–能力验证”为核心逻辑，整合计算机视觉、语音处理与大语言模型三大技术模块，构建面向课堂场景的多模态数据采集、融合、分析体系。通过引导学生完成系统架构设计、多模态融合算法实现、课堂感知效果验证等全流程操作，不仅帮助学生掌握多模态融合技术的核心原理与工程实现方法，更培养其从教育教学需求出发设计技术方案、以评价标准验证技术有效性的综合能力，助力其掌握课堂整体感知的核心方法。

本实验的设计与实施，既是对人工智能师范专业实验教学体系的补充与优化，也是推动人工智能技术与师范教育深度融合的重要实践。通过实验，学生能够深入理解智慧课堂的技术逻辑与教育价值，切实提升课堂整体感知能力，为未来从事智能教育产品研发、智慧课堂教学实践奠定坚实基础，同时助力培养适应新时代智能教育发展需求的复合型师资人才。

2. 多模态融合驱动智慧课堂感知系统的实验优势与教学价值

多模态融合驱动的智慧课堂感知系统实验，相较于传统单一模态的智能教育实验，具有显著的实验优势与贴合人工智能师范专业人才培养的教学价值，既兼顾工程实践的科学性与可操作性，又凸显师范教育的针对性与实用性，为人工智能师范本科生搭建了技术实践与教育应用衔接的重要桥梁。

在实验优势方面，传统课堂行为分析存在明显局限，其核心缺陷在于依赖人工编码的方式开展数据采集与分析，主观判断性强、误差较大，且存在样本量小、费时费力的问题；同时，传统分析所依赖的数据类型较为单一，仅能捕捉师生坐立、讲授、应答等外显行为，难以精准反映学习者的情绪波动、认知状态、生理变化等内隐特征，无法全面还原课堂教学的真实过程与内在规律，难以满足智能时代课堂教学分析的需求。

针对上述局限，基于多模态数据的课堂教学行为分析实验展现出显著优势。该实验以多模态融合为核心，有效突破了传统单一模态数据信息片面、感知维度有限的弊端，通过整合视频、语音、文本信号、人机交互等多类型数据，进行协同处理与深度融合，实现了智慧课堂教学状态感知的全流程闭环。这一模式不仅大幅提升了课堂感知结果的全面性与精准度，更能真实还原智慧课堂中教学行为与学生表现的内在关联，有效弥补了传统单一模态实验及传统课堂行为分析的不足。同时，实验流程贴合真实智能教育场景，设计科学、可操作性强且容错率高，契合人工智能师范本科生的认知规律与学习节奏，能够帮助学生逐步理解并掌握多模态融合的核心逻辑，降低多模态技术的学习门槛，兼顾实验的规范性与易实施性，为多模态数据在课堂教学行为分析中的实践应用奠定了基础[4]。

在教学价值方面，该实验精准对接人工智能师范专业的人才培养目标，将多模态融合技术与智慧课堂教学场景深度绑定，打破了“技术学习”与“教育应用”之间的壁垒，引导学生从教育教学实际需求出发，运用人工智能相关技术解决课堂感知中的实际问题，有效提升学生的工程实践能力、跨模态融合思维与场景适配能力。此外，实验过程中，学生需完整参与数据处理、模型应用、结果分析等全流程操作，既能巩固人工智能相关核心知识，又能培养其数据思维、逻辑分析与问题解决能力，为未来从事智能教育相关工作、推进教育信息化建设奠定坚实基础，助力培养适应新时代智能教育发展需求的复合型师范人才，同时也为人工智能师范专业实验教学体系的优化完善提供了重要的实践参考。

3. 多模态融合驱动智慧课堂感知系统的整体实验流程

本实验以智慧课堂真实教学场景为核心载体，围绕“语音–视觉”双模态融合展开，整体遵循“数据采集–语音转文本与时间戳提取–教学行为分类–视觉动作识别–双模态结果整合”的分步实施逻辑，兼顾实验的可操作性与师范专业的教学适配性，具体流程如图 1 所示。

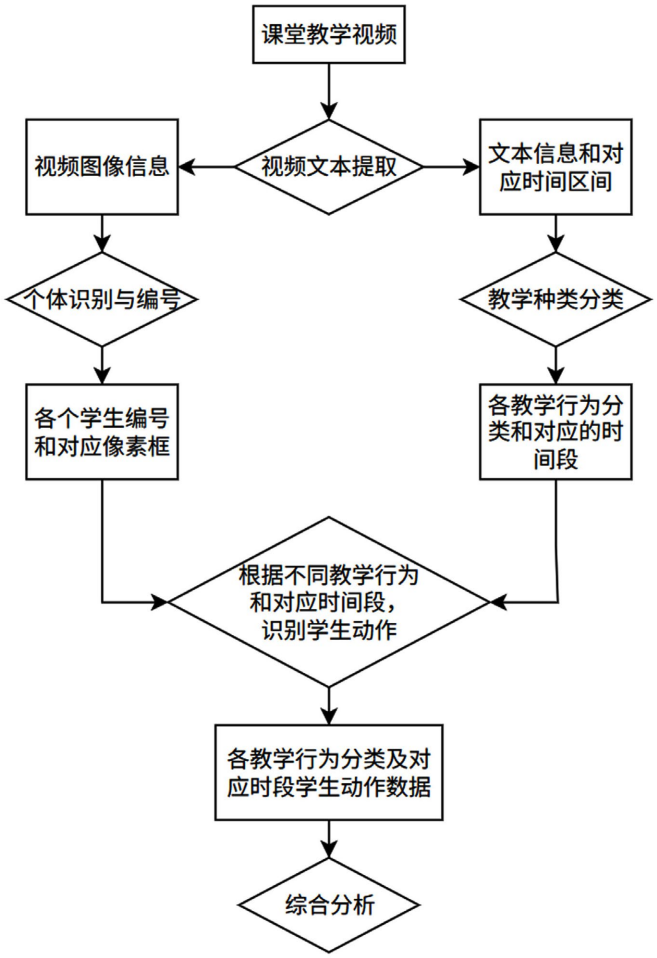

Figure 1. Overall flowchart
图 1. 整体流程图

首先，搭建智慧课堂双模态数据采集环境，采用高清摄像头录制完整课堂教学视频(获取视觉模态数据)、同步采集课堂语音信息(获取语音模态数据)，确保视频画面清晰、师生动作可识别，语音无杂音、语义可分辨，同步记录视频与语音的实时时间轴，为后续双模态时序对齐提供基础。

其次，引入大语言模型对提取的语音文本数据进行教学行为分类，结合人工智能师范专业的教学场景需求，预设“知识点讲解、课堂提问、小组讨论、练习巩固、课堂小结”等典型教学行为类别，将文本数据输入大语言模型，通过模型推理输出每段文本对应的教学行为类别，同步关联原有时间戳及对应语音、视频片段，形成“教学行为类别–时间段”的双模态标签数据。

同时，采用目标检测模型(YOLO 系列)对课堂视频流中的学生进行实时定位与唯一编号，结合课堂场景特点优化检测策略，精准识别每一位学生的人体区域并输出边界框坐标，为每位学生分配唯一标识编

号，确保编号与学生个体一一对应、全程可追溯。该定位编号结果将同步关联教学行为分类标签及时序信息，为后续学生动作识别、行为分析及“教学内容–学生表现”关联图谱构建提供精准的个体锚定基础，与前文视频推理、语音文本分析环节形成完整的数据关联闭环。

然后，基于前文教学行为分类数据及学生定位编号数据，针对不同教学行为类别对应的时间段，对各编号学生逐一进行动作识别，通过改进型 ViT 模型对学生人体区域特征进行推理判定，最终整合得到“学生行为类别–时间段–各编号学生动作”的关联数据，实现教学行为与学生个体动作的精准对应，为后续多模态数据融合分析奠定基础。

最后，基于上述整合得到的“学生行为类别–时间段–各编号学生动作”关联数据，调用提前依据课堂教学评估标准训练完成的大语言模型，开展课堂教学情况综合评估。该模型以预设的教学评估指标(如学生参与度、教学环节适配性、师生互动有效性等)为依据，对整合后的数据进行深度分析，结合教学行为类别与学生动作表现的对应关系，输出客观、全面的课堂教学评估结果，实现从数据采集、分析到教学评估的完整实验流程，为智慧课堂教学质量优化及师范教学示范提供数据支撑与决策参考。

4. 多模态融合驱动智慧课堂感知系统各实验具体流程

4.1. 课堂教学视频文本提取与教学种类分类

本实验环节聚焦语音模态的深度挖掘与语义解析，核心目标是完成智慧课堂教学内容的结构化切分与教学阶段精准划分。作为连接“语音文本转换”与“视觉模态分析”的关键枢纽，该环节的处理精度直接决定后续多模态融合分析的整体效果，是保障智慧课堂教学行为分析准确性的核心前提。

实验流程遵循“音频预处理–语音转写–语义解析–时序标注”的逻辑展开，具体操作如下：首先，对原始课堂视频进行音频轨剥离，筛选出清晰的课堂语音流，摒弃环境杂音、无关干扰音等无效音频片段，为后续转写提供高质量数据基础。随后，采用具备高准确率、强鲁棒性及多场景适配能力的 Whisper 语音识别模型[5]，对预处理后的整段课堂音频流进行全时段、高精度转写。该过程不仅实现了从语音信号到文本信息的精准转换，更核心的优势的是能够同步输出每一段转写文本对应的高精度时间戳，精确标记每一句语音的起始时刻、终止时刻及持续时长，确保转写文本与原始视频帧在时间轴上实现毫秒级严格对齐，为后续教学行为与视觉画面的联动分析奠定基础。

在语音转写完成后，引入大语言模型(LLM) [6]对转写文本进行深度语义理解与教学行为分类。结合人工智能师范专业的课堂教学场景特征，结合教学规律与实际课堂场景，将课堂教学内容系统归纳为五大核心教学阶段：知识点讲解、课堂提问、小组讨论、练习巩固、课堂小结。大语言模型依托其强大的上下文理解能力与语义分析能力，通过解析语句的语法结构、词汇语义、语用逻辑及课堂场景特征，实现教学阶段的精准判定——例如，通过识别“首先”“其次”“因此”等逻辑引导词，判定知识点讲解阶段；通过识别“谁来回答一下”“大家思考一下”等典型句式，定位课堂提问环节；通过识别“小组内交流”“互相讨论”等表述，划分小组讨论阶段。

本研究采用 API 调用方式实现大语言模型的轻量化接入，无需本地部署模型，通过 Python 代码完成实时文本输入与教学阶段输出，整体流程包括文本输入、prompt 规则约束、模型推理、结果结构化解析四部分。模型接入遵循统一输入输出标准，输入为语音转写后的文本段，输出为固定格式的教学阶段标签，保证系统稳定、可扩展、可复现。

最终，通过大语言模型的语义解析与分类，输出一份按时间轴有序排列的结构化时序标签，明确标注每个教学阶段的起止时间戳、对应教学行为类别及核心文本依据，形成完整的“时间戳–教学行为类别–核心文本”时序标签数据集。该数据集将为后续视觉模态的定向分析提供精确的时间窗口指引，实

现语音文本信息与视觉画面信息的精准联动，为智慧课堂教学行为的全方位、多维度分析提供有力支撑。

4.2. 学生个体识别与编号

本实验环节聚焦视觉模态下的目标检测与身份锚定任务，核心解决“复杂课堂场景中精准识别并持久化区分每位学生”的工程难题，是实现学生个体行为时序分析的核心基础。实验依托 4.1 环节构建的时序标签数据集，对不同教学阶段对应的视频片段实施逐帧或抽帧处理，为学生个体识别与追踪提供高质量的输入样本。

实验选用 Ultralytics YOLO26 作为核心目标检测模型[7] [8]，兼容 YOLOv5、YOLOv8、YOLO11 等主流版本，该模型作为 Ultralytics YOLO 系列的前沿迭代产品，以边缘优化为核心设计理念，引入多项关键创新：移除分布焦点损失(DFL)以简化导出流程、采用原生端到端无 NMS 推理消除后处理瓶颈，结合渐进损失平衡(ProgLoss)和小目标感知标签分配(STAL)提升训练稳定性与小目标检测精度，同时集成 MuSGD 优化器，实现受 LLM 启发的稳定收敛，有效解决传统注意力模型优化过程中的不稳定性问题。该模型兼具边缘部署便捷性、高检测精度与低推理延迟，CPU 和 Jetson 设备推理速度较前代提升显著，可高效适配课堂实时处理需求。

针对课堂场景的特殊性，对 YOLO26 模型进行针对性优化训练，重点强化对师生混杂、学生走动、部分遮挡等典型干扰因素的鲁棒性，依托模型自带的 STAL 机制，确保能够精准识别视频画面中每位学生的人体区域，有效过滤教师、背景等非目标对象的干扰。识别过程中，系统实时输出每个学生检测框的像素坐标(x1, y1, x2, y2)，如图 2 所示，精确定位学生在画面中的空间位置，为后续行为分析提供空间坐标支撑。该模型经优化后，在保持实时推理性能的同时，进一步提升了复杂场景下的检测精度，契合课堂视觉感知任务的核心需求。

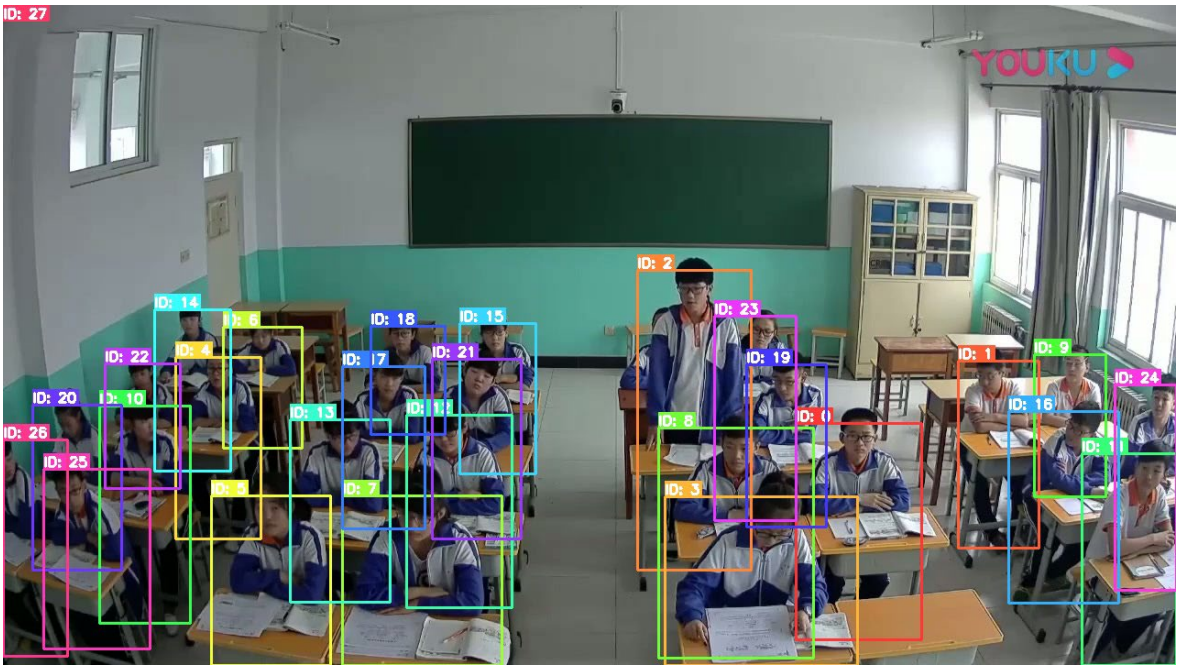

Figure 2. Sample example diagram of student recognition completed in the SAV dataset
图 2. SAV 数据集中完成学生识别的样本示例图

为实现学生个体的持久化区分，可采用“卡尔曼滤波 + IoU 匹配”的多帧关联算法[9] [10]，完成学

生目标的跨帧追踪与全局唯一编号分配。该算法结合 YOLO26-S 原生无 NMS 推理的低延迟特性，可实时预测并更新学生目标的运动状态，有效解决遮挡、视角变化、学生位置移动等场景下的编号漂移问题；同时依托 MuSGD 优化器带来的模型参数稳定性，进一步提升编号的一致性与可靠性，确保同一学生在不同时间段、不同拍摄角度，甚至部分遮挡的情况下，其全局唯一编号始终保持不变，为后续学生个体行为的时序关联分析提供坚实的身份锚定基础。相较于前代 YOLO 模型，YOLO26-S 的端到端设计简化了追踪算法的集成流程，降低了系统 latency，更适配课堂视频实时处理的应用场景。

4.3. 学生个体动作识别

本实验环节实现视觉模态的精细化动作识别，并与前序环节的语音模态数据深度融合，最终完成“谁–在什么阶段–做了什么”的全链路智慧课堂状态还原。实验首先利用 SAV (Student Action Video, SAV-Dataset)数据集对改进型视觉 Transformer (ViT)模型进行针对性的微调与迁移学习[11]，依托该数据集涵盖的 15 类学生课堂动作(如表 1 所示)、758 个不同教室的真实场景数据及多标签标注信息(含姿态、人–物交互等细分类别)，有效解决通用模型在课堂场景中对小目标、密集物体、遮挡场景识别精度不足的问题，显著提升模型在复杂课堂环境下的动作识别鲁棒性，契合实验对课堂行为识别的核心需求。

Table 1. Five major categories and fifteen subcategories of the SAV dataset
表 1. SAV 数据集的 5 个大类和 15 个小类

序号	动作大类	SAV 数据集动作标签
1	姿势动作	sit (坐姿)
2	姿势动作	stand (站姿)
3	视线动作	look forward (向前看)
4	视线动作	look sideways (四处看)
5	视线动作	read (阅读)
6	人–物交互	flip a book (翻书)
7	人–物交互	raise hand (举手)
8	人–物交互	take notes (记笔记)
9	人–物交互	clap (鼓掌)
10	人–物交互	hands down (放下手)
11	人–物交互	touch (触摸/持物)
12	身体运动动作	bend (弯腰)
13	身体运动动作	turn around (转身)
14	人–人交互	talk with others (与他人交谈)
15	人–人交互	answer questions (回答问题)

模型训练阶段，我们基于 MMAction2 框架(OpenMMLab 下一代视频理解工具箱) [12]，选用并改进了基于 ViT 的基线模型。该框架具备良好的通用性和扩展性，可高效适配 SAV 这类类 AVA 格式的数据集[13]——因 SAV 数据集与 AVA 数据集在动作标注规范、时空格式定义上高度相似，MMAction2 框架内置的类 AVA 数据集适配模块，能够直接兼容 SAV 数据集的输入格式与标注规则，无需额外进行大量格式适配开发。

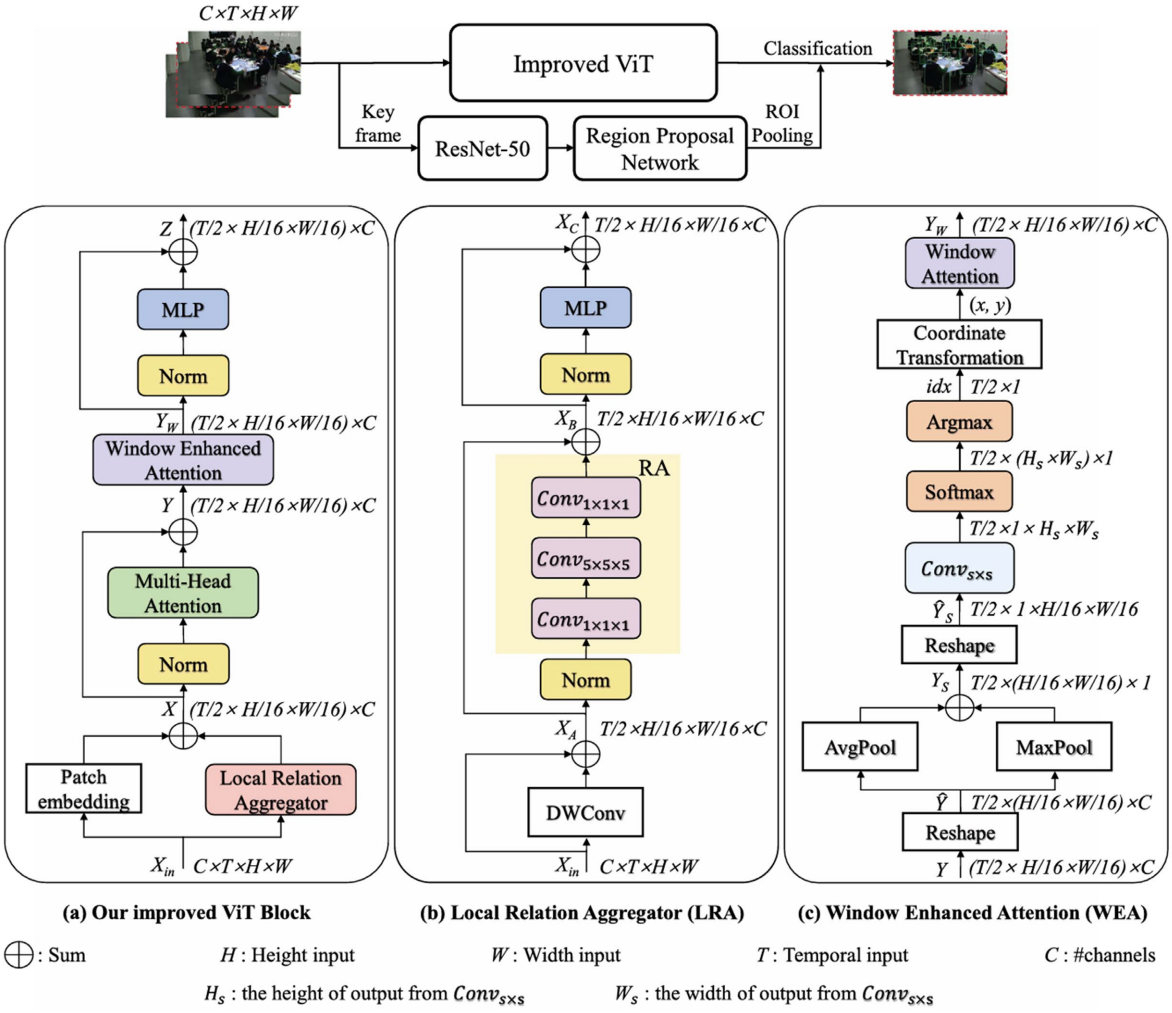

Figure 3. (a) Improved ViT module; (b) Local Relation Aggregator (LRA); (c) Window Enhanced Attention (WEA) [11]
图 3. (a) 改进后的 ViT 模块；(b) 局部关系聚合器(LRA)；(c) 窗口增强注意力(WEA) [11]

模型整体结构与计算流程均严格遵循文献[11]的设计，主要包括改进型 ViT 模块、LRA 局部特征增强模块与 WEA 关键区域注意力模块，如图 3 所示。

具体如下：

1) 改进型 ViT 模块推理流程

改进 ViT 模块将输入特征依次通过局部关系增强、多头注意力与窗口增强注意力计算，如图 3(a)所示，其核心推理过程如下：

$$X = \mathrm{PE}\left(X_{in}\right) + \mathrm{LRA}\left(X_{in}\right), \tag{1}$$

$$Y = \mathrm{MHA}\left(\mathrm{Norm}\left(X\right)\right) + X, \tag{2}$$

$$Z = \mathrm{MLP}\left(\mathrm{Norm}\left(\mathrm{WEA}\left(Y\right)\right)\right) + \mathrm{WEA}\left(Y\right). \tag{3}$$

该流程完成全局特征与局部关键特征的联合建模，使模型更适应课堂复杂场景。

2) 局部关系聚合器(LRA)

LRA 模块通过深度卷积与关系聚合实现局部特征增强，计算过程为：

$$X_a = \mathrm{DWConv}\left(X_{in}\right) + X_{in}, \tag{4}$$

$$X_b = \mathrm{RA}\left(\mathrm{Norm}\left(X_a\right)\right) + X_a, \tag{5}$$

$$X_c = \mathrm{MLP}\left(\mathrm{Norm}\left(X_b\right)\right) + X_b. \tag{6}$$

该模块能够增强模型对学生手部、肢体等小目标区域的特征感知能力。

3) 窗口增强注意力(WEA)

WEA 模块则通过响应图定位关键窗口，完成时空注意力增强：

$$Y_s = \mathrm{Sum}\left(\mathrm{AvgPool}\left(\hat{Y}\right), \mathrm{MaxPool}\left(\hat{Y}\right)\right), \tag{7}$$

$$idx = \mathrm{Argmax}\left(\mathrm{Softmax}\left(\mathrm{Conv}_{S\times S}\left(\hat{Y}_S\right)\right)\right). \tag{8}$$

随后通过坐标变换定位窗口位置：

$$\left(x, y\right) = \left(mod\left(idx, u\right), \left\lfloor \left(idx, u\right) \right\rfloor\right). \tag{9}$$

在对应窗口内执行时空注意力增强：

$$Y_W = \mathrm{attn}\left(Y\left(x : x + w, y : y + w\right)\right). \tag{10}$$

通过定位响应最强窗口，WEA 可有效聚焦书写、举手、阅读等关键动作区域。

模型训练完成后进入推理阶段：系统依据预设教学阶段时间窗口及学生唯一编号、边界框坐标，对视频流进行定向分析，设定每 30 帧(约 1 秒，适配常规视频帧率)进行一次动作识别，形成实验闭环。具体而言，系统锁定特定时间段内目标学生人体区域，每 30 帧提取一次关键帧的骨骼关键点及表观特征，输入经 SAV-Dataset 微调的改进型 ViT 模型推理判定。模型通过 LRA、WEA 模块，针对性应对 SAV 数据集典型挑战，借鉴其目标检测与多标签动作分类思路提升识别精度；采用“时序标签引导 + 个体编号锚定”模式，剔除背景干扰、聚焦目标学生行为。最终输出包含学生编号、教学阶段、时间点、动作类别的四维关联表格，格式参考 SAV-Dataset 标注规范，确保数据规范可追溯。该结果实现视觉模态动作识别与语音模态教学阶段分类的深度融合，构建“教学内容–学生表现”多模态关联图谱，充分发挥 SAV-Dataset 的数据与技术支撑作用。

4.4. 多模态融合分析

作为本实验的收尾与升华环节，本阶段核心目标是整合前文 4.1 语音语义分析、4.2 学生个体定位及 4.3 动作识别三大模块的成果，通过“预设判断标准 + 大语言模型(LLM)驱动”的智能分析模式，构建可解释、可落地的多维度数据关联体系，最终生成具备高决策价值的智慧课堂感知综合报告，实现实验全流程的闭环落地。

首先，结合师范专业教学评价规范、SAV-Dataset 行为标注逻辑及课堂教学规律，制定涵盖量化指标计算、行为模式分析、教学决策建议三大维度的结构化判断标准体系，明确课堂参与度、学生专注度等核心指标的计算方法，界定不同教学阶段的行为适配标准及对应的优化方向，将这些判断标准转化为大语言模型可理解的结构化训练指令，对大语言模型进行针对性微调训练，确保模型能够精准贴合智慧课堂分析场景，熟练运用预设标准解读多模态数据。

随后，系统将前三阶段产出的“时间戳–教学行为”时序标签集、“时间戳–学生编号–检测框像素坐标”定位数据集、“学生编号–动作类别”行为数据集，通过统一时空索引技术完成标准化处理与

关联融合，依托时间轴精准对齐机制，打破各模块数据壁垒，形成包含教学行为、学生检测框、个体动作的结构化多模态输入数据，输入至经判断标准训练后的大语言模型中。大语言模型依托其强大的上下文理解与推理能力，严格遵循预设判断标准，对融合数据进行智能化分析，不仅能自动量化整体课堂参与度、学生专注度等核心指标，精准统计不同教学阶段的互动数据，更能深度挖掘教学行为与学生反馈的联动关系，识别课堂中的行为模式差异、互动不均衡等问题，同时结合判断标准中的决策建议规则，生成贴合师范生教学实践的针对性优化建议。

最终，大语言模型沿用 4.1 节中既定的调用方法，输出完整的智慧课堂感知综合报告。该报告不仅为师范生提供直观、精准的课堂教学反馈，助力其科学优化教学节奏、灵活调整师生互动策略，夯实教学实践能力；同时为智慧教育相关算法的迭代优化提供可靠、可追溯的数据支撑，完整闭环多模态融合驱动的智慧课堂感知系统。此举充分发挥了各实验环节的实践价值与应用意义，成功实现了从“数据整合”到“智能决策”的核心升华，彰显了多模态技术在智慧课堂场景中的应用价值。

5. 讨论与反思

5.1. 真实课堂的复杂性考量与模型优化方向

本研究构建的多模态融合驱动智慧课堂感知系统，虽依托 SAV 数据集与主流算法实现了核心识别目标，但真实课堂的复杂性远超实验预设，仍有优化空间。真实课堂的动态性与不确定性主要体现在三方面：一是教学阶段非完全互斥，当前互斥标签难以刻画混合式教学行为；二是学生动作意图模糊，单一标签无法完整涵盖真实状态；三是随机场景干扰易降低识别精度。针对以上问题，后续研究需构建更灵活的识别模型，引入非互斥标签与时序事件检测模型适配混合教学场景，同时通过增加突发样本标注、数据增强等技术，提升模型抗干扰能力与场景适配性。

5.2. 系统应用的伦理考量与隐私保护

智慧课堂感知系统因需实时监测师生行为，必然涉及隐私与伦理问题，处理不当易引发隐私泄露、教育不公等隐患，因此需明确应用边界、建立完善防护机制，确保技术服务于教育教学并规避负面影响；其伦理核心为“以人为本”，明确系统仅作为辅助教学工具，杜绝将其数据作为师生评价唯一标准，坚守教育公平、尊重师生主体地位并提前获得知情同意；同时针对课堂敏感数据建立全流程隐私保护闭环，采集时遵循“最小必要”原则并脱敏，存储时加密授权，处理时优先本地运算并匿名化，使用仅限教学研究且实验结束后彻底销毁，同时通过隐形监测、不做动作价值判断、强化师生素养培养等方式，规避负面教育影响，实现技术赋能教育而非绑架教育。

6. 结论与展望

本文围绕人工智能师范本科生实验教学需求，设计并阐述了基于语音－视觉双模态融合驱动的智慧课堂感知系统实验方案，通过一系列分阶段实验的设计与实施，得出以下结论：该实验方案有效实现了智慧课堂教学阶段分类、学生个体识别与编号、学生动作时序识别的全流程闭环，依托 Whisper 模型、YOLO26 模型与大语言模型的协同作用，成功完成语音－视觉双模态数据的时序对齐与信息融合，能够精准呈现课堂教学行为与学生动作表现的关联关系，实验流程规范、可操作性强，既符合人工智能技术的应用逻辑，又贴合师范专业的教学场景需求。

同时，该实验不仅具备显著的技术优势，更能精准对接人工智能师范专业人才培养目标，有效弥补传统单一模态实验的不足，帮助学生巩固核心技术知识、提升工程实践能力与跨模态融合思维，为复合型智能教育师资人才的培养提供了有效的实践载体，也为人工智能师范专业实验教学体系的优化完善提

供了重要参考。

结合实验实施情况与智能教育领域的发展趋势，本文对后续实验优化与研究方向提出以下展望：在实验优化方面，可进一步丰富多模态数据类型，加入文本模态(如课件文本、学生作业反馈)与生理模态(如学生注意力生理指标)，构建多维度融合的智慧课堂感知体系，提升感知结果的全面性与精准度；同时，可优化模型训练策略，结合课堂场景的特殊性构建专属数据集，进一步降低模型识别误差、提升实验实用性。在教学应用方面，可将该实验方案与师范生教育实习相结合，引导学生将实验成果应用于真实课堂教学分析，助力其更好地掌握智能教育技术的应用方法，实现“实验实践–教学应用”的深度衔接。在技术拓展方面，可引入边缘计算技术，优化系统实时处理能力，实现智慧课堂感知结果的实时反馈，为教师教学决策提供及时支撑；同时，可探索大语言模型与多模态模型的深度协同，提升系统对课堂教学场景的语义理解与行为预测能力，推动智慧课堂感知系统向智能化、个性化方向发展，为新时代智能教育的高质量发展注入新活力，也为人工智能师范专业人才培养提供更贴合行业需求的实践路径。

基金项目

天津师范大学 2023 年教学改革研究项目《人工智能(师范)专业教育实践课程建设与人才培养体系研究》(项目编号：JG01223085)。

参考文献

[1] Zhou, J., Ran, F., Li, G., Peng, J., Li, K. and Wang, Z. (2022) Classroom Learning Status Assessment Based on Deep Learning. *Mathematical Problems in Engineering*, **2022**, Article ID: 7049458. https://doi.org/10.1155/2022/7049458

[2] Jia, L., Sun, H., Jiang, J. and Yang, X. (2025) High-Quality Classroom Dialogue Automatic Analysis System. *Applied Sciences*, **15**, Article 1613. https://doi.org/10.3390/app15031613

[3] Liu, Q., Jiang, X. and Jiang, R. (2025) Classroom Behavior Recognition Using Computer Vision: A Systematic Review. *Sensors*, **25**, Article 373. https://doi.org/10.3390/s25020373

[4] 张乐乐, 顾小清. 多模态数据支持的课堂教学行为分析模型与实践框架[J]. 开放教育研究, 2022, 28(6): 101-110.

[5] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2023) Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the* 40*th International Conference on Machine Learning* (*ICML*'23), **202**, 28492-28518.

[6] Chu, Z., Wang, S., Xie, J., *et al*. (2025) LLM Agents for Education: Advances and Applications. arXiv: 2503.11733.

[7] Jocher, G., Qiu, J. and Chaurasia, A. (2023) Ultralytics YOLO (Version 8.0.0) [Computer Software]. https://github.com/ultralytics/ultralytics

[8] Sapkota, R. and Karkee, M. (2025) Ultralytics YOLO Evolution: An Overview of YOLO26, YOLO11, YOLOv8 and YOLOv5 Object Detectors for Computer Vision and Pattern Recognition. arXiv: 2510.09653.

[9] Wojke, N., Bewley, A. and Paulus, D. (2017) Simple Online and Realtime Tracking with a Deep Association Metric. 2017 *IEEE International Conference on Image Processing* (*ICIP*), Beijing, 17-20 September 2017, 3645-3649. https://doi.org/10.1109/icip.2017.8296962

[10] Arioka, K. and Sawada, Y. (2023) Improved Kalman Filter and Matching Strategy for Multi-Object Tracking System. 2023 62*nd Annual Conference of the Society of Instrument and Control Engineers* (*SICE*), Tsu, 6-9 September 2023, 772-777. https://doi.org/10.23919/sice59929.2023.10354112

[11] Tan, Z., Gao, C., Qin, A., Chen, R., Song, T., Yang, F., *et al*. (2025) Towards Student Actions in Classroom Scenes: New Dataset and Baseline. *IEEE Transactions on Multimedia*, **27**, 6831-6844. https://doi.org/10.1109/tmm.2025.3590899

[12] Contributors, M. (2020) OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2

[13] Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., *et al*. (2018) AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 6047-6056. https://doi.org/10.1109/cvpr.2018.00633