2026 届研究生硕士学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：71255901010

華東師範大學

East China Normal University

硕士学位论文

MASTER'S DISSERTATION

基于课堂录像的教师风格画像分析系统

院系：	**计算机科学与技术学院**
专业：	**电子信息**
研究方向：	**计算机技术**
指导教师：	**陈蕾 副教授**
学位申请人：	**陶瑞鑫**

2026 年 2 月 20 日

Dissertation for Master's Degree in 2026

University Code: 10269

Student ID: 71255901010

EAST CHINA NORMAL UNIVERSITY

A Teacher Style Profiling Analysis System Based on Classroom Video

Department:	School of Computer Science and Technology
Major:	Computer Application Technology
Research Direction:	Computer Vision
Supervisor:	Prof. Li Si
Candidate:	Zhang San

March, 2026

摘　要

在教育数字化转型的浪潮中，海量课堂录像数据亟待被有效利用以赋能教学。教师教学风格是影响课堂质量的关键因素，但传统评价方法主观性强、反馈滞后，难以满足智慧教育环境下对客观、实时、可量化课堂反馈的需求。为此，本研究设计并实现了一个基于多模态深度学习的教师教学风格画像分析系统，旨在提供客观、精细、可解释的智能风格画像分析。

现有课堂分析技术存在单模态视频或音频难以全面刻画教学风格和风格识别无法提供决策依据和特征贡献度等问题。针对上述挑战，本研究提出了 **SHAPE（Semantic Hierarchical Attention Profiling Engine，语义层次化注意力画像引擎）**，通过语义驱动分段、层次化教学意图识别和跨模态注意力机制实现特征的自适应融合与风格的精准画像。具体包括：数据分段策略优化方面，提出语义驱动的话语分段策略，通过依存句法分析和话语边界检测，保持教学话语的语义完整性，使教学意图识别准确率大幅度提升；音频模态方面，不仅将音频用于语音情绪识别，在课堂场景下进行微调，使用自动语音识别（ASR）技术将音频转化为文本模态，为意图识别打下基础；文本模态方面，引入基于 BERT 的层次化细粒度对话行为识别（Hierarchical Dialogue Act Recognition），采用两层分类架构（粗分类 4 类 + 细分类 10 类），将单层分类扩展为双层 10 类细粒度分类，更有效地捕捉不同教学风格的特征性语言模式；视觉模态方面，使用身份识别算法实现稳定的教师身份追踪，并采用时空图卷积网络对骨骼序列进行时序建模，相比单帧动作识别准确率较大提升；智能融合与解释方面，设计的 SHAPE 通过跨模态注意力机制自适应地融合视觉、音频、文本特征，并结合注意力权重与 SHAP 可解释性分析，提升模型决策依据的可追溯性。

在自建的教师风格数据集（209 个样本，7 类风格）上，SHAPE 在风格识别任务中取得了 **92.5%** 的准确率，显著优于单一模态方法和简单融合方法。消融实

验进一步证实，语义驱动分段策略使风格识别准确率提升 **7.6 个百分点**（McNemar $\chi^2 = 4.00$，$p < 0.05$），验证了这些改进的有效性。

同时构建出一套教师课堂画像系统，将上述算法的结果进行可视化。本系统能够生成直观、可追溯的教师风格画像（风格雷达图、模态贡献度分析、典型片段回放），为教师风格认知和教学研究提供了科学、客观、精细化的数据支撑。

关键词： 教师教学风格；多模态学习分析；跨模态注意力；深度学习

ABSTRACT

In the wave of educational digitalization, massive volumes of classroom video data urgently need to be effectively leveraged to empower teaching. Teaching style is a key determinant of classroom quality; however, traditional evaluation methods are highly subjective and suffer from delayed feedback, making them inadequate for the demand for objective, real-time, and quantifiable classroom feedback in smart education environments. To address this, the present study designs and implements a teacher teaching-style profiling system based on multimodal deep learning, aimed at delivering objective, fine-grained, and interpretable intelligent style analysis.

Existing classroom analysis techniques are constrained by two core limitations: single-modality video or audio is insufficient to comprehensively characterize teaching styles, and style recognition models fail to provide decision rationales or feature contribution insights. To address these challenges, this study proposes **SHAPE (Semantic Hierarchical Attention Profiling Engine)**, which achieves adaptive feature fusion and accurate style profiling through semantic-driven segmentation, hierarchical teaching intent recognition, and a cross-modal attention mechanism. The system incorporates the following specific contributions: (1) *Data segmentation*: a semantic-driven utterance segmentation strategy is proposed, which preserves the semantic integrity of teaching utterances through dependency parsing and discourse boundary detection, substantially improving teaching intent recognition accuracy; (2) *Audio modality*: audio is employed not only for speech emotion recognition fine-tuned to classroom scenarios, but also converted into text via Automatic Speech Recognition (ASR), laying the foundation for intent recognition; (3) *Text modality*: BERT-based Hierarchical Dialogue Act Recognition (H-DAR) is introduced with a two-level classification architecture (4 coarse classes + 10 fine classes), extending single-level classification to fine-grained dual-level recognition and more effectively capturing the characteristic linguistic patterns of different teaching styles; (4) *Visual modal-*

ity: identity recognition algorithms are employed to achieve stable teacher tracking, and Spatial-Temporal Graph Convolutional Networks (ST-GCN) perform temporal modeling of skeleton sequences, substantially outperforming single-frame action recognition; (5) *Fusion and interpretability*: SHAPE adaptively fuses visual, audio, and text features via a cross-modal attention mechanism, and combines attention weights with SHAP interpretability analysis to enhance the traceability of model decision rationale.

On a self-constructed teacher style dataset (209 samples, 7 style categories), SHAPE achieves an accuracy of **92.5%** on the style recognition task, significantly outperforming single-modality methods and simple fusion approaches. Ablation experiments further confirm that the semantic-driven segmentation strategy improves style recognition accuracy by **7.6 percentage points** (McNemar $\chi^2 = 4.00$, $p < 0.05$), validating the effectiveness of these improvements.

A teacher classroom profiling system is also constructed to visualize the outputs of the above algorithms. The system generates intuitive and traceable teacher style profiles—including style radar charts, modality contribution analysis, and representative segment playback—providing scientific, objective, and fine-grained data support for teacher style awareness and pedagogical research.

Keywords: *Teacher Teaching Style; Multimodal Learning Analytics; Cross-modal Attention; Deep Learning*

目　录

图目录

表目录

第一章 绪 论

1.1 研究背景及意义

在教育现代化与数字化转型的浪潮中，课堂教学正从” 资源配置与教学辅助” 阶段迈向” 智能评价与数据驱动决策” 阶段。众多学校与教育管理部门通过录播系统、教学平台、课堂监控设备等手段，积累了大量课堂录像、音频记录和教学日志。然而，这些过程性数据往往仅用于教学回看或行政存档，缺乏对教学特征刻画与教师风格认知的持续支撑。

传统课堂评价方式——包括听课记录、专家评估、学生问卷及访谈等——在主观性、时效性和覆盖面方面均存在显著局限，难以满足智慧教育环境下对” 客观、实时、可量化” 课堂反馈的需求。尤其在 K-12 阶段，讲授式课堂在知识传授与课堂组织中仍占据主导地位，如何通过数据化方式刻画教师风格、反映教学特征，成为实现课堂精细化分析的重要课题。

在此背景下，教师教学风格作为连接课堂行为与教学效果的重要中介变量，逐渐受到学界与实践界的广泛关注。教学风格通常包含教师在语言表达、课堂互动、非言语行为、情感表达等多维度上的稳定特征，直接影响学生的学习动机与课堂氛围。如果能够通过多模态数据（视频、音频、文本）构建教师风格的可解释画像模型，不仅可以为教师提供客观的风格认知，也能够为教学研究、教师培训及教育决策提供科学依据。

此外，课堂对于教师风格还具有明显的动态性与情境依赖性：不同学段、学科、教学内容下，适宜的教学风格存在差异；教师的风格亦会随教龄增长与理念更新而变化。这种复杂性进一步提高了人工观察与主观评价的难度，也凸显了以人工智能技术实现风格建模与反馈的必要性。

因此，本研究以课堂视频为核心输入，融合语音、文本等多模态数据，重点探讨教师教学风格的量化映射机制与智能识别体系的实现路径。在理论层面，本研究旨在丰富教育人工智能领域关于多模态课堂分析与教师画像建模的研究体系；在应用层面，则期望构建一个能够自动化识别教师行为、提取语音语义特征、生成

可解释风格画像的系统，以促进教师风格认知与教学研究。

1.2 国内外研究现状

教师教学风格识别技术的发展经历了从理论抽象到数据驱动、从单一模态到多模态融合的演进过程。本节将从教师风格理论基础、课堂多模态分析技术和融合方法三个维度梳理相关研究进展。

1.2.1 教学风格：从理论分类到计算建模

教师教学风格是指教师在长期教学实践中形成的、相对稳定的教学行为模式和个性化特征。教学风格的研究经历了从理论抽象到量化分析、从人工观察到自动识别的演进过程，逐步形成了” 理论分类 → 行为编码 → 技术增强 → 数据驱动 → 智能识别” 的发展脉络。

教学风格的量化研究起源于 20 世纪 60 年代的课堂互动分析。Bellack 等人（1966）提出的课堂语言游戏理论，将课堂互动分解为” 引发——回应——反应——评价” 四阶段循环，奠定了互动分析的早期理论基础。

Flanders 系统提出的互动分析系统（FIAS[1]，Flanders Interaction Analysis System）是最早成熟的课堂行为编码工具，通过 10 类编码对课堂互动进行量化记录：教师言语包括接纳情感、表扬鼓励、接受学生想法、提问、讲授、给予指导、批评维权共 7 类；学生言语包括回应和主动发言 2 类；沉默或混乱 1 类。FIAS 建立了” 教师话语比例”” 学生参与度”” 间接影响指数” 等量化指标，开创了课堂行为的结构化分析范式。

S-T 分析法（Student-Teacher Interaction Analysis）是另一类经典课堂分析工具，它将课堂行为简化为教师行为（T）与学生行为（S）两类，通过绘制 S-T 时序图与 Rt-Ch 图，可区分练习型、讲授型、对话型、混合型等课堂结构，从整体上判断课堂互动的主导模式。

这些早期方法共同推动了课堂行为从定性描述走向可观测、可量化的科学分析，为教学风格的量化研究提供了重要的方法支撑。

随着教育心理学和认知科学的持续发展，教学风格的研究重心逐渐从课堂行

为量化转向理论层面的系统分类，研究者开始从更宏观的视角界定教学风格的核心维度与类型。Grasha（1996）[2] 提出了经典的五分类模型，将教师教学风格划分为专家型（核心强调知识传授的专业性与学科内容的深度挖掘）、权威型（重点突出课堂秩序的维护与教学规范的执行）、示范型（通过自身教学行为示范，引导学生模仿学习）、促进型（注重激发学生主动性，支持学生自主探索与合作学习）、委托型（充分下放学习自主权，最大化发挥学生的自主管理与学习能力）。该模型明确了教师在知识控制、课堂结构搭建与师生互动关系中的核心差异，成为目前国际上应用最广泛、影响力最深远的教学风格分类框架之一。

Pianta 等人（2008）[3] 开发的 CLASS 评价工具（Classroom Assessment Scoring System），则突破了传统教学风格分类的局限，从” 情感支持”” 课堂组织”” 教学支持” 三个核心维度评估课堂教学质量，通过标准化的观察量表与评分标准，将教学风格的质性特征转化为可量化的评估指标，首次系统建立了教学风格与教学效果之间的实证关联，为教学风格的量化评估提供了新的思路与工具。

在国内研究领域，学者钟启泉（2001）立足中国基础教育实践情境，从教学理念、教学策略、师生互动关系等核心维度，提出了适配中国课堂情境的教学风格分类体系，明确区分了” 传递-接受型”” 引导-发现型”” 自主-探究型” 等典型教学风格类型，弥补了国外理论模型在本土教学情境中的适配性不足，丰富了教学风格的理论研究体系。

进入 21 世纪，信息技术的快速发展为课堂分析方法的革新注入了新动力，推动教学风格量化研究向” 技术融合” 方向拓展。顾小清等（2007）[4] 立足 Flanders 互动分析系统的核心框架，针对多媒体教学环境的特殊性与需求，通过新增师生与技术之间的互动维度，设计开发出 ITIAS（Information Technology-based Interaction Analysis System，基于信息技术的互动分析编码系统）。与 FIAS[1] 相比，ITIAS 在经典的” 师-生” 二元互动分析基础上，新增了” 教师操作技术”” 学生使用技术”” 技术呈现内容” 等专属编码类别，构建起” 师-生-技” 三元互动的课堂分析框架，有效适配了交互白板、投影仪、平板电脑等技术工具广泛应用的新型课堂环境，该系统也因其较强的本土适配性，在国内中小学信息化教学研究中得到广泛应用与推广。

步入 2010 年代，随着教育大数据技术与学习分析（Learning Analytics）领域

的兴起，数据驱动的教师画像（Teacher Profiling）成为教学风格研究的新热点与核心方向。胡小勇等（2018）从教研数据采集、分类整理以及多源数据有效关联等关键角度，系统阐释了数据驱动视角下教师画像 [5] 构建的实施路径与可行性。该框架核心强调多源课堂与教研数据的融合应用，涵盖课堂录像、教案文本、学生作业、考试成绩等多元数据类型，通过数据挖掘技术对各类数据进行分析处理，构建起涵盖教学风格、教学能力、教研水平等维度的教师画像标签体系。与此同时，Worsley & Blikstein（2013）提出的多模态学习分析（Multimodal Learning Analytics, MMLA）框架，进一步推动了教师课堂行为的多维度、精细化刻画，该框架通过整合视频、音频、眼动、手势等多源多模态数据，打破了单一数据类型的局限，构建了更加全面、立体的课堂行为分析体系。这一时期，教育数据挖掘（Educational Data Mining）领域开始广泛尝试运用聚类分析、关联规则挖掘、序列模式挖掘等技术方法，从海量课堂数据中自动发现、提炼教学行为模式，这一转变标志着教学风格研究正式从传统的” 理论分类” 向现代的” 数据建模” 完成范式跨越，推动研究走向更加科学、精准、智能化的新阶段。

近年来，深度学习技术的突破为教师风格的自动识别提供了新的可能。卷积神经网络（CNN）在课堂视频分析中的应用，使得教师动作识别（如走动、板书、手势、指向等）无需人工设计特征即可从原始像素中学习高层语义表示。循环神经网络（RNN）和长短期记忆网络（LSTM）被用于建模课堂互动的时序依赖，捕捉” 提问-等待-回应-反馈” 等序列模式。Transformer 架构及其注意力机制在语音识别和语义理解中的成功，使得教师话语的自动转写和教学意图识别成为可能。预训练语言模型（如 BERT）在课堂对话分析中的应用，能够识别教师话语中的提问、指令、讲解、反馈等对话行为。多模态融合技术的发展，使得研究者能够综合视频、音频、文本等多源信息构建教师风格的整体画像。

1.2.2 多模态分析技术

课堂教学是一个复杂的多模态交互过程，涉及教师的语言表达、肢体动作、情感状态等多个维度。随着人工智能技术的发展，语音识别、文本理解、视频分析等技术在课堂场景中的应用日益深入，为教师风格的自动识别提供了技术基础。

语音语义识别技术

语音识别技术经历了从统计模型到深度学习、从监督学习到自监督学习的发展历程。早期的语音识别主要基于声学特征提取和统计建模。在特征提取方面，梅尔频率倒谱系数（MFCC）是最广泛使用的特征表示，通过模拟人耳对不同频率声音的感知特性，将音频信号转换为若干维的特征向量。此外，滤波器组特征（FBANK）、感知线性预测系数（PLP）等也被广泛应用。在建模方面，隐马尔可夫模型（HMM）结合高斯混合模型（GMM）构成了传统语音识别的主流框架，通过统计建模捕捉语音信号的时序特性和状态转移规律。这些方法在特定场景下取得了一定效果，但依赖大量的人工特征工程和复杂的系统构建。

深度学习的兴起带来了语音识别的革命性变化。Hannun 等人（2014）[6] 提出的 DeepSpeech 系统采用循环神经网络（RNN）实现了端到端的语音识别，直接从原始音频波形学习到文本的映射，无需人工设计中间特征表示。该系统采用连接时序分类（CTC，Connectionist Temporal Classification）作为损失函数，解决了输入序列与输出序列长度不一致的对齐问题，开启了语音识别的深度学习时代。Chan 等人（2016）提出的 Listen, Attend and Spell（LAS）模型引入了注意力机制（Attention Mechanism），通过编码器-解码器架构实现了更加灵活的序列到序列建模，显著提升了识别准确率。

自监督学习的兴起进一步突破了对大量标注数据的依赖。Baevski 等人（2020）[7] 提出的 Wav2Vec 2.0 通过自监督对比学习从无标注音频中学习通用的声学表征。该方法首先使用卷积神经网络提取音频的局部特征，然后通过 Transformer 网络建模长程依赖，最后通过对比学习目标（contrastive learning）学习区分真实语音片段和负样本。Wav2Vec 2.0 在仅使用少量标注数据的情况下，在多种下游任务（语音识别、情感识别、说话人识别等）上取得了显著性能提升，成为语音处理领域的重要里程碑。HuBERT（Hidden-Unit BERT）进一步改进了自监督学习策略，通过聚类-预测的方式学习离散的声学单元，实现了更好的语音表征。

端到端语音识别模型的发展达到了新的高度。Radford 等人（2022）[8] 提出的 Whisper 模型通过在 68 万小时多语言多任务数据上进行弱监督训练，实现了接近人类水平的语音识别能力。Whisper 采用 Transformer 编码器-解码器架构，支持多

语言识别、语音翻译、语言识别、语音活动检测等多个任务，在真实场景的鲁棒性上表现出色。针对课堂环境的特殊性，CPT-Boosted Wav2Vec2.0（2024）通过持续预训练（Continued Pretraining）在课堂域数据上进行适配，进一步提升了在噪声环境下的鲁棒性，有效应对了课堂中的学生讨论声、椅子移动声、空调噪声等干扰。

在语音情感识别方面，传统方法主要基于韵律特征（pitch、energy、duration）和频谱特征（MFCC）进行建模。深度学习方法通过端到端的网络直接从原始音频学习情感表示。3D 卷积神经网络（3D-CNN）能够同时捕捉频谱的时间和频率维度的特征，循环神经网络（RNN/LSTM）则擅长建模情感的时序演化。最新的研究将 Wav2Vec 2.0 等预训练模型应用于情感识别，通过在情感数据集上进行微调（fine-tuning），在自然对话和课堂场景中取得了优异的性能。

说话人分离与识别技术在多人课堂场景中尤为重要。x-vector 系统通过时延神经网络（TDNN）提取说话人嵌入向量，能够在变长语音中稳定地识别说话人身份。ECAPA-TDNN（Emphasized Channel Attention, Propagation and Aggregation TDNN）进一步引入了通道注意力机制和多层特征聚合，显著提升了说话人识别的准确率。这些技术使得在课堂录像中自动区分教师和学生的语音、分析师生话轮转换模式成为可能。

文本语义识别技术

文本语义识别技术经历了从浅层表征到深层语义理解的演进过程。早期的课堂对话分析主要依赖关键词匹配和规则方法。通过预定义的词表和句式模板，研究者可以识别教师话语的类型，例如包含” 为什么”” 怎么” 等疑问词的句子被标记为提问，包含” 请”” 大家” 等词的句子被标记为指令。TF-IDF（Term Frequency-Inverse Document Frequency）方法通过统计词频和逆文档频率，提取文档的关键词特征。词袋模型（Bag of Words）和 N-gram 模型则通过统计词语或词语序列的出现频率进行文本分类。这些方法实现简单，但难以捕捉语言的深层语义、上下文依赖和语序信息。

词嵌入技术（Word Embedding）的出现标志着文本表征的重要进步。Mikolov 等人（2013）[9] 提出的 Word2Vec 通过神经网络学习词语的分布式表示，将词语映射到连续的低维向量空间。Word2Vec 包括两种训练方式：CBOW（Continuous Bag of

Words）通过上下文词预测中心词，Skip-gram 通过中心词预测上下文词。Pennington 等人（2014）提出的 GloVe（Global Vectors）结合了全局矩阵分解和局部上下文窗口方法，通过共现矩阵的对数双线性回归学习词向量。Bojanowski 等人（2017）提出的 FastText 进一步引入了子词（subword）信息，通过字符级 N-gram 增强了对低频词和词形变化的建模能力。这些词嵌入方法使得语义相近的词语在向量空间中距离更近，为后续的文本分析任务奠定了基础。

序列建模技术的发展使得文本的上下文理解成为可能。循环神经网络（RNN）通过隐状态的循环连接建模序列的时序依赖，但在长序列中存在梯度消失问题。长短期记忆网络（LSTM，Long Short-Term Memory）通过引入门控机制（输入门、遗忘门、输出门）解决了长程依赖建模的难题。门控循环单元（GRU，Gated Recurrent Unit）进一步简化了 LSTM 的结构，在保持性能的同时降低了计算复杂度。双向 LSTM（BiLSTM）通过同时建模前向和后向的上下文信息，能够更全面地理解句子的语义。这些序列模型被广泛应用于文本分类、命名实体识别、关系抽取等任务。

注意力机制（Attention Mechanism）的引入进一步提升了序列建模能力。Bahdanau 等人（2015）在机器翻译任务中首次引入注意力机制，使得模型能够在生成每个输出词时动态地关注输入序列的不同部分。自注意力机制（Self-Attention）通过计算序列中每个元素与其他元素的关联程度，捕捉长程依赖和全局信息。Vaswani 等人（2017）[10] 提出的 Transformer 架构完全基于自注意力机制，抛弃了循环结构，通过多头注意力（Multi-Head Attention）和位置编码（Positional Encoding）实现了高效的并行计算和强大的表示能力。Transformer 成为自然语言处理领域的基础架构，催生了后续的预训练语言模型革命。

预训练语言模型的兴起带来了自然语言理解的突破。Devlin 等人（2018）[11] 提出的 BERT（Bidirectional Encoder Representations from Transformers）通过在大规模语料上进行掩码语言模型（Masked Language Model，MLM）和下一句预测（Next Sentence Prediction，NSP）的预训练，学习到了丰富的语言知识。BERT 采用双向 Transformer 编码器，能够同时利用左侧和右侧的上下文信息。RoBERTa（Robustly Optimized BERT Pretraining Approach）通过移除 NSP 任务、增大批大小、延长训练时间等优化策略，进一步提升了模型性能。ALBERT（A Lite BERT）通过参数

共享和因子分解降低了模型参数量，实现了轻量化部署。ELECTRA（Efficiently Learning an Encoder that Classifies Token Replacements Accurately）通过判别式预训练任务替代生成式任务，提升了训练效率。DeBERTa（Decoding-enhanced BERT with Disentangled Attention）通过解耦的注意力机制和增强的掩码解码器进一步提升了性能。这些预训练模型在文本分类、命名实体识别、问答系统、情感分析等任务上取得了突破性进展。

大语言模型（Large Language Models, LLMs）的出现进一步拓展了文本理解的边界。OpenAI 的 GPT 系列（GPT-1、GPT-2、GPT-3、GPT-4）通过自回归语言建模在海量文本上进行预训练，展现出强大的文本生成和少样本学习（few-shot learning）能力。Google 的 T5（Text-to-Text Transfer Transformer）将所有 NLP 任务统一为文本到文本的格式，实现了任务间的知识迁移。Meta 的 LLaMA 系列通过优化的训练策略在相对较小的参数规模下达到了与 GPT-3 相当的性能。ChatGPT 和 GPT-4 等对话式大语言模型通过指令微调（instruction tuning）和人类反馈强化学习（RLHF），展现出强大的对话能力、推理能力和知识整合能力。这些大语言模型在课堂对话分析中的应用，使得教师话语的深层语义理解、教学逻辑链分析、知识点提取、概念关系构建等高级任务成为可能。

视频与行为识别技术

视频行为识别技术经历了从手工特征到端到端深度学习的发展历程。早期的方法主要基于手工设计的特征描述子。方向梯度直方图（HOG，Histogram of Oriented Gradients）通过统计图像局部区域的梯度方向分布描述物体外观，光流直方图（HOF，Histogram of Optical Flow）通过统计光流的方向分布描述运动模式，运动边界直方图（MBH，Motion Boundary Histogram）通过计算光流的梯度来描述运动边界。时空兴趣点（STIP，Spatio-Temporal Interest Points）通过检测视频中显著的局部时空结构进行特征提取。密集轨迹（Dense Trajectories）方法通过在密集采样的兴趣点上跟踪轨迹，并提取轨迹周围的 HOG、HOF、MBH 特征，在动作识别任务上取得了很好的效果。

深度学习的引入极大地推动了视频分析技术的发展。早期的研究尝试将 2D 卷积神经网络应用于视频分析。Karpathy 等人（2014）探索了多种 2D CNN 在视频上

的应用方式，包括单帧建模、晚期融合、早期融合、慢融合等策略。AlexNet、VGG、ResNet 等在图像分类任务上取得成功的网络结构被迁移到视频领域，通过在视频数据集（如 UCF-101、HMDB-51）上进行微调实现了一定的性能提升。

Simonyan & Zisserman（2014）[12] 提出的双流网络（Two-Stream Network）是视频分析的重要里程碑。该方法通过两条并行的卷积神经网络分别处理 RGB 外观信息和光流运动信息，空间流网络（Spatial Stream）从单帧 RGB 图像中学习外观特征，时间流网络（Temporal Stream）从堆叠的光流图像中学习运动特征，最后融合两路特征进行动作识别。Wang 等人（2016）提出的时间分段网络（TSN，Temporal Segment Networks）在双流网络基础上引入了稀疏采样策略，将长视频分为若干段，在每段中随机采样一帧，通过分段共识函数（segment consensus function）聚合多段的预测结果，实现了长时序建模。

3D 卷积神经网络的引入使得时空特征的联合学习成为可能。Tran 等人（2015）[13] 提出的 C3D（3D Convolutional Networks）通过 3×3×3 的 3D 卷积核同时在空间和时间维度进行特征提取，学习到了通用的视频表征。Carreira & Zisserman（2017）[14] 提出的 I3D（Inflated 3D ConvNet）将在 ImageNet 上预训练的 2D 卷积网络” 膨胀” 为 3D 卷积网络，通过在 Kinetics 大规模视频数据集上进行预训练，实现了更好的时空建模能力。Feichtenhofer 等人（2019）[15] 提出的 SlowFast 网络通过双路径设计，Slow 路径以低帧率捕捉语义信息，Fast 路径以高帧率捕捉运动信息，两路径通过横向连接进行信息交互，实现了效率和性能的平衡。

基于骨骼序列的图卷积网络（GCN）方法提供了一种更高效的视频分析方案。OpenPose（2017）通过自底向上的方法实现了实时的多人姿态估计，提取人体的关键点坐标（如头部、肩膀、肘部、手腕、臀部、膝盖、脚踝等）。MediaPipe（2019）[16] 进一步提供了轻量化的姿态估计解决方案，能够在移动设备上实时运行。Yan 等人（2018）[17] 提出的 ST-GCN（Spatial Temporal Graph Convolutional Networks）将人体骨骼序列建模为时空图结构，节点表示关节点，边表示关节间的连接关系（骨骼连接和时间连接），通过图卷积捕捉关节间的空间依赖和时间演化。相比于基于 RGB 的方法，骨骼序列表征不仅计算效率更高（特征维度从百万级降至百级），而且天然具有抗遮挡和隐私保护的优势，特别适合教育场景的应用。Shi 等人（2020）

提出的 MS-G3D（Multi-Scale Graph Convolutional Networks）通过多尺度时空图卷积和解耦的时空建模进一步提升了骨骼序列动作识别的性能。

Transformer 架构的引入进一步提升了视频理解能力。Dosovitskiy 等人（2021）[18] 提出的 ViT（Vision Transformer）将图像分割为 patch 序列，通过 Transformer 编码器进行建模，在图像分类任务上取得了与 CNN 相当甚至更好的性能。Liu 等人（2021）[19] 提出的 Video Swin Transformer 将窗口注意力机制扩展到视频领域，通过局部窗口和跨窗口的注意力计算，在保持高效计算的同时建模长程时空依赖。Bertasius 等人（2021）提出的 TimeSformer 通过分解的时空注意力机制（先空间注意力再时间注意力），实现了高效的视频理解。这些基于 Transformer 的方法在多个视频理解基准上刷新了性能记录，注意力机制的可解释性也为理解模型决策提供了重要途径。

目标检测技术在课堂场景分析中发挥着重要作用。YOLO（You Only Look Once）系列通过单阶段检测实现了实时的物体定位和分类，能够在课堂视频中检测教师、学生、黑板、课桌等物体。Faster R-CNN 通过区域提议网络（RPN）和 Fast R-CNN 的结合，实现了高精度的目标检测。姿态估计技术的发展使得对教师肢体语言的细粒度分析成为可能。AlphaPose 通过自顶向下的方法实现了鲁棒的多人姿态估计，HRNet（High-Resolution Network）通过保持高分辨率表示提升了关键点定位的精度。

在教育场景的具体应用中，Gupta 等人（2021）使用姿态估计结合 LSTM 时序建模识别教师的典型动作（如讲解、板书、走动、指向等）。最新的 MM-TBA 数据集（2024）收集了超过 300 位教师的 4,839 个教学视频片段，涵盖讲解、板书、走动、互动、手势、指向等 6 类典型教学动作，为教师行为识别算法的训练和验证提供了标准化的基准。该数据集发表于 Nature Scientific Data 期刊，包含丰富的标注信息（动作类别、时间戳、边界框、姿态关键点等），成为该领域重要的公开资源。YOLOv8 结合可变形大核注意力（DLKA）机制（2024）能够在复杂场景下准确识别小目标（如教师的手势细节、学生的举手动作），显著提升了课堂行为检测的鲁棒性。ClassMind 系统（2025）[20] 采用多模态大语言模型（LLM）作为核心分析引擎，通过 AVA-Align 流水线实现了对课堂视频的长上下文推理和时序定位，能够

自动生成教师的等待时长、师生对话平衡、学生参与度等量化指标。EduSpatioNet（2025）将 YOLOv8 目标检测与时空图神经网络（GNN）结合，通过建模师生的空间关系和时序交互，实现了教师行为识别与专家评估的高一致性。这些研究表明，深度学习技术已经能够在真实课堂环境中实现高精度、可解释的行为识别。

1.2.3 多模态融合方法

单一模态的分析存在固有的局限性：仅分析语音无法捕捉肢体语言的丰富性，仅分析视频则忽略了语义内容的重要性，仅分析文本则缺失了情感和非言语信息。多模态融合通过整合不同模态的互补信息，成为提升分析性能的关键。多模态融合方法经历了从浅层拼接到深层交互、从固定权重到自适应学习、从黑盒模型到可解释分析的演进过程。

早期融合策略：特征拼接与决策加权

早期的多模态融合研究主要采用特征级拼接（Early Fusion）或决策级融合（Late Fusion）的策略。特征级拼接是最直接的融合方式，将不同模态的特征向量简单拼接后输入统一的分类器。例如，将视频特征 $F_v \in \mathbb{R}^{d_v}$、音频特征 $F_a \in \mathbb{R}^{d_a}$、文本特征 $F_t \in \mathbb{R}^{d_t}$ 拼接为联合特征 $F_{concat} = [F_v; F_a; F_t] \in \mathbb{R}^{d_v+d_a+d_t}$，然后通过全连接层或 SVM 进行分类。这种方法实现简单，但存在明显的问题：不同模态的特征维度和尺度差异大，高维模态会主导融合结果；模态间的语义关联被忽略，例如教师”指向黑板”（视觉）与”请看这个公式”（文本）的协同语义关系无法被捕捉。

决策级融合（Late Fusion）则采用分而治之的策略，为每个模态训练独立的分类器，然后对各模态的预测结果进行加权融合。常见的融合方式包括平均融合、加权平均、投票机制等。例如，加权平均融合的预测结果为 $P_{final} = w_v P_v + w_a P_a + w_t P_t$，其中 P_v, P_a, P_t 是各模态的预测概率，w_v, w_a, w_t 是权重系数（通常手工设置或通过验证集优化）。这种方法允许各模态独立建模，但权重系数是全局固定的，无法根据样本内容自适应调整。

混合融合（Hybrid Fusion）尝试结合早期融合和晚期融合的优势，在网络的中间层进行特征融合。Karpathy 等人（2014）在视频分类中探索了多种融合时机：单帧融合、后期融合、早期融合、混合融合等。Ngiam 等人（2011）提出的多模态深

度玻尔兹曼机（Multimodal DBM）通过共享隐层表示实现模态融合。然而，这些方法仍然依赖于固定的网络结构，缺乏对模态交互的动态建模。

Worsley & Blikstein（2013）首次系统性地提出了" 多模态学习分析"（MMLA，Multimodal Learning Analytics）的概念框架，倡导整合视频、音频、眼动、手势、生理信号等多源数据进行学习过程分析。该框架强调了多模态数据的时间同步、特征对齐、联合建模等技术挑战，为后续的多模态融合研究提供了理论指导。然而，早期的 MMLA 研究多采用简单的特征拼接或结果加权，未能充分挖掘模态间的深层交互关系。

注意力机制的引入：模态间的动态交互

注意力机制（Attention Mechanism）的引入为多模态融合带来了革命性的变化。Bahdanau 等人（2015）在神经机器翻译任务中首次引入注意力机制，使得解码器能够在生成每个目标词时动态地关注源序列的不同部分。这一思想很快被拓展到多模态学习中：不同模态可以通过注意力机制相互" 查询"，动态地提取相关信息。

交叉注意力（Cross-Attention）是多模态交互的核心机制。给定两个模态的特征表示 F_i 和 F_j，交叉注意力通过以下步骤计算模态 j 对模态 i 的增强表示：线性投影将特征投影到 Query、Key、Value 空间：

$$Q_i = F_i W_Q, \quad K_j = F_j W_K, \quad V_j = F_j W_V \tag{1.1}$$

通过 Query 和 Key 的相似度计算注意力权重：

$$\alpha_{i \to j} = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d_k}}\right) \tag{1.2}$$

根据权重聚合 Value：

$$\widetilde{F}_i^{(j)} = \alpha_{i \to j} V_j \tag{1.3}$$

这一机制使得模态 i 能够根据自身的内容（Query）动态地从模态 j 中提取相关信息（Value），实现了样本自适应的模态交互。

Vaswani 等人（2017）[10] 提出的 Transformer 架构将自注意力机制发展到了新

的高度。Transformer 通过多头注意力（Multi-Head Attention）并行计算多组 Query-Key-Value 投影，捕捉不同子空间的语义关联。位置编码（Positional Encoding）的引入使得 Transformer 能够建模序列的顺序信息。Transformer 的成功催生了一系列多模态预训练模型。

多模态预训练模型在大规模图文对数据上进行预训练，学习到了视觉和语言的对齐表示。ViLBERT（2019）采用双流架构，分别对图像和文本进行编码，然后通过 co-attention 层进行跨模态交互。LXMERT（2019）进一步引入了三种类型的编码器：目标关系编码器（对象间的空间关系）、语言编码器（文本语义）、跨模态编码器（视觉-语言交互）。UNITER（2020）和 VILLA（2020）通过统一的 Transformer 编码器联合建模图像和文本，采用掩码语言建模（MLM）、掩码区域建模（MRM）、图文匹配（ITM）等预训练任务学习跨模态表示。

对比学习为多模态对齐提供了新的范式。Radford 等人（2021）提出的 CLIP（Contrastive Language-Image Pre-training）通过对比学习在 4 亿图文对上进行预训练，学习到了强大的视觉-语言对齐能力。CLIP 的核心思想是最大化匹配图文对的相似度，同时最小化不匹配图文对的相似度。训练后的模型能够将图像和文本映射到统一的嵌入空间，实现零样本图像分类、图像检索等任务。ALIGN（2021）通过在更大规模的噪声图文对（18 亿）上训练，进一步提升了对齐能力。这些对比学习方法为多模态融合提供了强大的预训练基础。

统一多模态 Transformer：从双流到单流

多模态 Transformer 架构经历了从双流到单流的演进。双流架构（如 ViLBERT、LXMERT）为每个模态设计独立的编码器，然后通过跨模态交互层进行融合。这种设计保留了各模态的特定表示，但计算开销较大，且模态间的交互深度有限。

单流架构将所有模态的 token 统一输入到一个 Transformer 编码器中，通过自注意力机制同时建模模态内和模态间的依赖。Kim 等人（2021）提出的 ViLT（Vision-and-Language Transformer）是单流架构的代表，将图像 patch 和文本 token 拼接后输入 Transformer，通过自注意力机制实现深层的跨模态交互。ViLT 的优势在于：（1）简化了网络结构，减少了参数量；（2）通过端到端训练实现了更深层的模态融合；

Li 等人（2021）提出的 ALBEF（Align Before Fuse）引入了 momentum distillation

策略，通过教师模型指导学生模型学习更鲁棒的跨模态表示。Li 等人（2022）提出的 BLIP（Bootstrapping Language-Image Pre-training）通过 caption 和 filter 两个模块迭代优化，从噪声网络数据中学习高质量的图文对齐。BLIP 在图像描述、视觉问答、图像-文本检索等任务上取得了显著提升。

多模态大模型的出现进一步拓展了多模态融合的能力。Flamingo（2022）通过在冻结的大语言模型（LLM）中插入视觉条件的 cross-attention 层，实现了少样本视觉-语言学习。BLIP-2（2023）通过轻量化的 Q-Former 桥接冻结的视觉编码器和大语言模型，在保持高性能的同时大幅降低了训练成本。GPT-4V、Gemini 等多模态大模型展现出强大的视觉理解、推理和生成能力，标志着多模态融合进入了大模型时代。

可解释性：理解模型的决策依据

随着多模态深度学习模型在教育场景中的应用日益深入，可解释性（Explainability）成为关键需求。教育工作者需要理解模型的决策依据，而不仅仅是接受一个黑盒的预测结果。注意力权重可视化是最直观的解释方法，通过可视化跨模态注意力矩阵 $\alpha_{i\to j}$，可以展示不同模态之间的交互模式。例如，在教师风格识别中，如果音频模态对视觉模态的注意力权重较高，说明模型认为” 语音韵律” 与” 肢体动作” 之间存在强关联，这可能对应” 情感表达型” 教师的特征。

SHAP 值（SHapley Additive exPlanations[21]）提供了更严格的特征归因方法。注意力机制与 SHAP 值的互补性在于：注意力权重反映了模型内部的信息流动（哪些模态/特征被关注），SHAP 值是从博弈论角度给出的特征边际贡献（哪些特征影响了决策）。结合两者可以提供更全面的模型解释。例如，如果某个特征的注意力权重高但 SHAP 值低，说明该特征虽然被模型关注，但对最终预测的贡献不大；反之，如果某个特征的注意力权重低但 SHAP 值高，说明该特征虽然不被显式关注，但在决策中起到了关键作用。

1.3 研究目标与内容

本研究旨在构建一个基于课堂录像的教师风格画像分析系统，实现教学风格的量化建模、可解释映射与即时反馈。系统目标包括三个层面：

（1）建立多模态融合的教师风格分析框架，实现视频、音频与文本数据的协同建模；

（2）构建基于可解释特征的教师风格分类模型，支持风格画像与反馈；

（3）验证系统在真实课堂场景中的可行性与有效性，为教育评价提供数据支撑。

在当前课堂评价体系中，教师的课堂风格和行为特征是影响教学质量的重要因素。然而，传统评价方式（学生问卷、人工观课）普遍存在主观性高、反馈滞后、覆盖面窄等缺陷。为实现上述研究目标，我们将研究内容分为以下四个方面：

（1）构建教师风格映射模型：结合教育学理论与课堂实地观察，定义七类具有区分力的教学风格（理论讲授型、耐心细致型、启发引导型、题目驱动型、互动导向型、逻辑推导型、情感表达型），设计规则驱动与可解释机器学习结合的风格映射机制，实现多模态特征到风格标签的映射。

（2）设计非言语行为识别模型：利用时空图卷积网络对骨骼序列进行时序建模识别教师典型动作、空间分布与互动行为，并通过课堂场景数据集进行训练与验证。

（3）设计语音语义特征提取模块：采用基于 Transformer 的语音识别与情绪分析模型，提取语义特征（提问结构、关键词、逻辑连接词）与情绪特征（语调、语速、情感倾向）。

（4）设计风格映射与可视化机制：将行为与语言特征融合后，构建风格分类器及可视化模块，生成雷达图、得分分布、典型片段等可解释结果，支持教师风格认知与教学研究。

1.4 论文组织结构

本论文围绕” 基于课堂录像的教师风格画像分析系统” 这一主题展开，全文共分为五章，结构安排如下：

第一章绪论。本章阐述研究的背景与意义，分析传统课堂评价的局限性与智慧教育的发展需求，提出基于多模态数据实现教师教学风格建模的研究动机。同时，综述国内外相关研究现状，归纳多模态课堂分析、教师行为分析、语音语义识别与

视频动作识别等方向的研究进展，明确本研究的目标与内容，最后概述论文的整体结构与研究逻辑。

第二章理论基础与相关研究。本章从教育学与计算机科学的交叉视角，系统梳理教师教学风格的相关理论，包括教学风格的定义、分类及核心特征；分析课堂行为与语言特征的关联规律。在技术层面，介绍视频行为识别、音频识别与语音情绪分析、文本语义建模等多模态分析技术的基本原理与关键方法，为后续系统设计提供理论支撑。

第三章研究方法与总体设计。本章阐述研究的总体思路与框架结构，介绍多模态数据的采集与预处理流程，构建教师风格映射模型的设计思路与算法机制。重点描述行为特征与语音语义特征的融合方法、可解释风格分类机制的构建以及教师风格画像与反馈机制的总体设计思路，明确系统功能模块与技术路线。

第四章教师风格画像分析系统设计与实现。本章在前期研究与实验结果的基础上，介绍教师风格画像分析系统的设计与实现。内容包括系统总体架构、风格映射与画像生成模块、多模态特征可视化、风格雷达图及典型片段展示等。进一步阐述风格画像可视化与可解释性分析模块的设计理念，并展示系统的运行效果与应用场景，分析系统不足与优化方向。

第五章总结与展望。本章总结论文的主要研究成果，回顾系统的构建思路、实验结果与研究创新，分析研究中存在的问题与局限，最后对未来研究方向进行展望，包括在更大规模数据集上的模型验证、跨学科融合的应用拓展以及教学智能反馈机制的持续优化。

第二章　相关概念及研究

2.1　教师教学风格

教师教学风格（Teaching Style）是教育心理学与教学研究中一个重要而复杂的概念，反映教师在长期教学实践中形成的相对稳定的教学倾向、行为模式与交互特征。教学风格不仅体现教师在课堂中的教学理念与行为策略，也直接影响学生的学习动机、课堂氛围及教学效果。因此，教学风格的识别与建模是实现课堂智能分析与教学评价的重要理论基础。

早期的研究往往基于教学行为特征的分类。研究者在课堂观察与行为分析的基础上，将教师风格划分为讲授型、启发型、探究型、合作型、演示型等类型。例如，讲授型教师倾向于结构化知识讲解和板书展示；启发型教师注重提问、引导与学生参与；探究型教师侧重问题解决与任务驱动。这类划分便于将教学风格与具体课堂行为进行对应分析。

2.1.1　教师教学风格的量化测量

教学风格的量化研究起源于 20 世纪 60 年代的课堂互动分析。Flanders 提出的互动分析系统（FIAS[1]，Flanders Interaction Analysis System）是最早、最具代表性课堂行为编码工具，通过 10 类编码对课堂互动进行量化记录。FIAS 将课堂互动分为三大类：

（1）教师言语行为（7 类编码）

- **间接影响**（Indirect Influence）：
- 编码 1：接纳情感（Accepts Feeling）—— 接受并澄清学生的情感态度
- 编码 2：表扬鼓励（Praises or Encourages）—— 对学生行为给予正向反馈
- 编码 3：接受学生想法（Accepts or Uses Ideas of Students）—— 采纳学生观点并延展

- 编码 4：提问（Asks Questions）—— 向学生提出问题以引发思考
- **直接影响**（Direct Influence）：
- 编码 5：讲授（Lecturing）—— 陈述事实、观点或程序
- 编码 6：给予指导（Giving Directions）—— 发布指令或命令
- 编码 7：批评或维权（Criticizing or Justifying Authority）—— 批评学生行为或辩护教师权威

（2）学生言语行为（2 类编码）

- 编码 8：学生回应（Student Talk - Response）—— 回答教师提问
- 编码 9：学生主动发言（Student Talk - Initiation）—— 学生自发言语

（3）沉默或混乱

编码 10：沉默或混乱（Silence or Confusion）—— 可辨识的沉默或无法理解的混乱

FIAS 量化指标体系

基于 10 类编码，FIAS 建立了一套量化指标来描述教学风格：

1. **教师话语比例（Teacher Talk Ratio, TTR）**：

$$\mathrm{TTR} = \frac{N_{\text{teacher}}}{N_{\text{total}}} = \frac{N_1 + N_2 + \cdots + N_7}{N_1 + N_2 + \cdots + N_{10}}$$

其中，N_i 是编码 i 出现的次数，N_{total} 是总编码数。典型值：讲授型教师 TTR > 0.70，互动型教师 TTR < 0.60。

2. **间接影响比率（Indirect/Direct Ratio, I/D）**：

$$\text{I/D Ratio} = \frac{N_1 + N_2 + N_3 + N_4}{N_5 + N_6 + N_7}$$

该指标衡量教师是否倾向于间接引导（提问、鼓励）还是直接讲授。I/D > 1.0 表示间接影响占主导（启发型），I/D < 0.5 表示直接讲授占主导（传统型）。

3. **学生参与度（Student Participation Index, SPI）**：

$$\mathrm{SPI} = \frac{N_8 + N_9}{N_{\text{total}}} \times 100\%$$

典型值：讲授型课堂 SPI < 20%，互动型课堂 SPI > 40%。

4. **扩展学生想法比率（Extended Student Idea Ratio）**：

$$\mathrm{ESI} = \frac{N_3}{N_4} = \frac{\text{接受学生想法次数}}{\text{提问次数}}$$

ESI > 0.3 表示教师善于采纳并延展学生观点，体现启发引导型风格。

S-T 分析法（Student-Teacher Interaction Analysis）

S-T 分析法（Student-Teacher Interaction Analysis，简称 S-T 分析法）是一种简化型课堂教学互动定量分析方法，由日本教育学者藤田英典等人提出，旨在通过对课堂中" 教师行为" 与" 学生行为" 的二元划分，客观刻画课堂互动结构与教学模式。

与传统的弗兰德斯互动分析系统（FIAS[1]）等多维度编码体系不同，S-T 分析法采用极简二元分类，仅将课堂行为划分为两类：

T（Teacher）行为：教师讲授、板书、演示、提问、指导、评价等由教师主导的教学行为；

S（Student）行为：学生应答、思考、讨论、练习、操作等由学生发生的学习行为。

其核心分析思路为：以固定时间间隔对课堂教学过程进行连续采样，逐一刻画 T/S 行为序列，绘制 S-T 时序图，并计算两类关键量化指标：

Rt（教师行为占有率）：教师行为占整堂课的比例；

Ch（行为转换率）：课堂中 T 与 S 行为相互转换的频繁程度。

依据 Rt 与 Ch 的数值组合，可将课堂划分为练习型、讲授型、对话型、混合型四种典型教学模式，从而实现对课堂互动结构的量化判断与横向比较。

S-T 分析法的优势在于编码规则简单、主观性低、易操作、可重复，能够有效降低课堂观察的复杂度，适用于各类学科课堂教学互动的实证分析与教学评价。

CLASS 评价工具：从行为编码到多维评分 [3]

Pianta 等人（2008）[3] 开发的 CLASS 评价工具（Classroom Assessment Scoring System）代表了教学风格评价的重要进步。CLASS 不再采用逐秒编码的方式，而是通过标准化观察量表从三个维度评估教学质量：

CLASS 三维评价体系

1. **情感支持（Emotional Support）**：

   - 积极氛围（Positive Climate）：教师对学生的情感温暖、尊重和享受
   - 消极氛围（Negative Climate，反向计分）：愤怒、讽刺、严厉
   - 教师敏感性（Teacher Sensitivity）：对学生需求的觉察和回应
   - 尊重学生观点（Regard for Student Perspectives）：学生自主性和领导力

2. **课堂组织（Classroom Organization）**：

   - 行为管理（Behavior Management）：清晰的期望和有效的行为矫正
   - 生产力（Productivity）：时间管理和课堂流程的流畅性
   - 教学学习形式（Instructional Learning Formats）：活动的多样性和参与度

3. **教学支持（Instructional Support）**：

   - 概念发展（Concept Development）：分析、综合、创造性思维
   - 反馈质量（Quality of Feedback）：扩展性反馈和脚手架支持
   - 语言建模（Language Modeling）：开放性问题、对话和词汇丰富性

CLASS 评分机制

每个维度采用 7 点量表（1-7 分）进行评分，其中：

- 1–2 分：低质量（Low）
- 3–5 分：中等质量（Mid）
- 6–7 分：高质量（High）

最终的教学风格可以通过三维得分的组合来表征：

$$\text{Style Vector} = (\text{ES}, \text{CO}, \text{IS}) \in [1,7]^3$$

2.1.2　教学风格的理论分类

逐渐出现了基于教学情感与交互特征的分类。研究者关注教师情感表达、语音语调、肢体语言等非言语特征，将教学风格分为理性逻辑型、情感表达型、互动导向型、稳健控制型等类别。这类分类强调教师在课堂氛围营造与人际互动中的差异特征，为后续多模态风格识别提供了可操作的维度参考。

Grasha 教学风格量表（Teaching Style Inventory, TSI）

Grasha（1996）[2] 提出了著名的五类教学风格模型，将教师划分为：

1. **专家型（Expert）**：强调知识传授与学科深度，以教师为知识权威
2. **权威型（Formal Authority）**：强调课堂秩序、规范与结构化教学
3. **示范型（Personal Model）**：通过自身行为示范引导学生学习
4. **引导型（Facilitator）**：注重学生自主探索与问题解决
5. **委派型（Delegator）**：最大化学生自主权，教师作为顾问角色

Grasha 开发了教学风格量表（TSI）来量化测量这五种风格 [2]。TSI 包含 40 个题项，每个风格 8 个题项，采用 5 点 Likert 量表（1= 非常不同意，5= 非常同意）。例如：

- **专家型题项**：”我希望学生将我视为某一领域的专家”（I want students to perceive me as an expert in the field）
- **促进型题项**：”我更多地扮演课堂活动的设计者而非讲授者”（I design classroom activities more than lecture）

风格得分计算：

$$S_{\text{Expert}} = \frac{1}{8} \sum_{i=1}^{8} R_i^{\text{Expert}}$$

其中，R_i^{Expert} 是第 i 个专家型题项的评分（1-5）。类似地计算其他四个维度的得分。

风格分类决策规则：

Grasha 提出了基于得分阈值的分类规则：

- 如果 $S_k \geq 4.0$，则风格 k 为" 主导风格"（Dominant）
- 如果 $3.0 \leq S_k < 4.0$，则风格 k 为" 中等倾向"（Moderate）
- 如果 $S_k < 3.0$，则风格 k 为" 低倾向"（Low）

大多数教师表现为**混合风格**，例如：

$$\text{Style Profile} = \{\text{Expert}: 4.2,\ \text{Authority}: 3.8,\ \text{Personal Model}: 3.5,\ \text{Facilitator}: 2.8,\ \text{Delegator}: 2.3\}$$

这表示教师以专家型为主导，辅以权威型和示范型特征。

信息技术的发展推动了课堂分析方法的革新。顾小清等（2007）[4] 基于 Flanders 互动分析系统 [1]，针对多媒体教学环境的特点，设计出了 ITIAS（Information Technology-based Interaction Analysis System，基于信息技术的互动分析编码系统）。

ITIAS 的" 师-生-技" 三元互动模型

ITIAS 在传统的师-生互动之外，增设学生思考编码，分离沉寂与混乱，形成了 15 类编码：

教师行为（7 类）：1-7：保留 FIAS 的原有编码

学生行为（4 类）：

- 8：学生操作技术（Student Operating Technology）
- 9：学生回应
- 10：学生主动发言
- 11：学生协作讨论（Student Collaborative Discussion）

技术呈现（3 类）：

- 12：技术呈现内容（Technology Presenting Content）

- 13：技术支持互动（Technology Supporting Interaction）
- 14：技术辅助评价（Technology Assisting Assessment）

其他：15：沉默或混乱

技术整合度指标（Technology Integration Index, TII）：

$$\text{TII} = \frac{N_8 + N_{12} + N_{13} + N_{14}}{N_{\text{total}}} \times 100\%$$

TII > 30% 表示技术深度整合，15% < TII < 30% 为中度整合，TII < 15% 为低度整合。

技术-教学协同指标：

$$\text{T-I Synergy} = \frac{N_{12\to4} + N_{13\to8} + N_{14\to9}}{N_{12} + N_{13} + N_{14}}$$

其中，$N_{12\to4}$ 表示” 技术呈现内容” 后紧接” 教师提问” 的转移次数。高协同值（> 0.5）表示技术工具与教学策略有机结合。

随着教育大数据技术和学习分析（Learning Analytics）的兴起，数据驱动的教师画像（Teacher Profiling）成为新的研究方向。**胡小勇等（2018）的教师画像 [5] 框架**

胡小勇等从教研数据采集、分类以及有效关联等角度，提出了数据驱动下的教师画像实施框架：

聚类与风格建模

使用无监督学习方法（如 K-means、层次聚类）对教师进行分组：

$$\text{Clustering: } T_1, T_2, \ldots, T_N \to C_1, C_2, \ldots, C_K$$

其中，T_i 是第 i 个教师的特征向量，C_k 是第 k 个聚类（风格类别）。

聚类质量评估：

- **轮廓系数（Silhouette Coefficient）**：

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

其中，$a(i)$ 是样本 i 到同类其他样本的平均距离，$b(i)$ 是样本 i 到最近异类样本的平均距离。$s(i) \in [-1, 1]$，越接近 1 表示聚类质量越好。

- **Davies-Bouldin 指数（DB Index）**：

$$\mathrm{DB} = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d\left(c_i, c_j\right)}\right)$$

其中，σ_i 是簇 i 内样本的平均距离，$d\left(c_i, c_j\right)$ 是簇中心间距离。DB 指数越小表示聚类越紧凑且分离。

画像生成与反馈

为每个教师生成多维画像：

$$\mathrm{Profile}\left(T_i\right) = \{\mathrm{Style}: C_k, \mathrm{Features}: \mathbf{f}_i, \mathrm{Percentile}: P_i, \mathrm{Improvement}: \Delta_i\}$$

其中：

- C_k：所属风格类别
- $\mathbf{f}_i$：特征向量（如提问频率 =12 次/45 分钟，走动时长 =8 分钟）
- P_i：在同类型教师中的百分位排名
- Δ_i：与历史数据对比的变化趋势

使用有监督学习方法训练风格分类器：

$$P\left(y = k \mid \mathbf{x}\right) = \mathrm{softmax}\left(W_k^T\mathbf{x} + b_k\right)$$

其中，$\mathbf{x}$ 是教师的特征向量，y 是风格标签，W_k 和 b_k 是模型参数。

常用的分类算法包括：

- **支持向量机（SVM）**：通过核函数映射到高维空间，寻找最优分类超平面

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

其中，$K(\mathbf{x}_i, \mathbf{x})$ 是核函数（如 RBF 核：$K(\mathbf{x}_i, \mathbf{x}) = exp\left(-\gamma \| \mathbf{x}_i - \mathbf{x} \|^2\right)$）

- **随机森林（Random Forest）**：通过集成多棵决策树提升泛化能力

$$\hat{y} = \text{mode}\{h_1(\mathbf{x}), h_2(\mathbf{x}), \ldots, h_T(\mathbf{x})\}$$

其中，$h_t(\mathbf{x})$ 是第 t 棵决策树的预测

- **深度神经网络（DNN）**：通过多层非线性变换学习复杂特征

$$\mathbf{h}^{(l+1)} = \sigma\left(W^{(l)}\mathbf{h}^{(l)} + \mathbf{b}^{(l)}\right)$$

其中，σ 是激活函数（如 ReLU、Tanh），l 是层索引

2.1.3　教师教学风格的核心特征

教学风格的多样性既反映教师个体差异，也体现学科特征与教学情境的差别。不同风格类型在课堂管理、知识呈现与情感互动中的优势互补，通常可从语言特征、非言语行为特征、课堂互动特征、教学组织特征四个方面加以刻画。为本研究后续的风格映射模型提供了理论支撑。

1. 语言特征。教师的语言风格是教学风格最直接的表现形式。语速、语调、停顿频率、情绪色彩以及关键词使用频率等要素均能反映教师的认知风格与教学策略。例如，理论讲授型教师更体现为注重核心名词的精准解释与技术发展演化的系统讲解；启发引导型教师则更频繁使用疑问句与引导性表达。通过语音识别与文本语义分析，可量化这些差异。
2. 非言语行为特征。教师的姿态、手势、面部表情、移动路径等非言语行为能够反映其课堂控制力与情感表达倾向。行为活跃度较高的教师往往具备较强的课堂调动能力，而动作单一或空间范围受限的教师则偏向传统讲授型风格。
3. 课堂互动特征。互动频率与话轮转换比例是衡量教师风格的重要指标。互动导向型教师倾向于与学生进行多轮交流，学生语音占比高；而讲授型教师课

堂中教师话语主导，学生参与度低。通过语音分离与对话检测技术，可以量化这类互动特征。

4. 教学组织特征。包括教学环节的结构化程度、任务驱动频率及教学节奏控制等方面。逻辑推导型教师在知识结构组织与时间控制上更为严谨；情感表达型教师则在课堂氛围与参与感营造方面更突出。

综上所述，教师教学风格不仅是个体教学理念的体现，更是多模态行为与语言特征在特定教学情境中的综合表达。对这些核心特征的深入分析，为本研究提供了明确的理论基础与分析维度。

2.2 教育场景中的多模态分析技术

教育场景中的多模态分析（Multimodal Analysis in Education）是近年来教育人工智能领域的重要研究方向。课堂活动是一种典型的多模态交互过程，教师的语言、动作、姿态、表情、语调及课堂互动等因素共同构成了复杂的多维信号体系。随着计算机视觉、语音识别与自然语言处理技术的快速发展，多模态学习分析（Multimodal Learning Analytics, MMLA）逐渐成为理解教学行为与学习过程的重要手段。本节将从视频、音频与文本三个角度，介绍课堂场景中常用的多模态分析技术原理与方法。

2.2.1 视频行为识别的原理与关键技术

视频行为识别（Video Action Recognition）旨在从连续视频帧序列中自动识别特定的人体动作或交互行为，是多模态课堂分析的核心技术之一。在课堂环境中，教师的讲解、走动、板书、手势、指示与互动等行为都能通过视频识别得到结构化表示，从而为教学风格建模提供行为层面的量化依据。

（1）传统方法：基于手工特征的视频分析

早期的视频行为识别主要基于手工设计的特征描述子。方向梯度直方图（HOG，Histogram of Oriented Gradients）通过统计图像局部区域的梯度方向分布描述物体外观，光流直方图（HOF，Histogram of Optical Flow）通过统计光流的方向分布描述运动模式，运动边界直方图（MBH，Motion Boundary Histogram）通过计算光流

的梯度来描述运动边界。时空兴趣点（STIP，Spatio-Temporal Interest Points）通过检测视频中显著的局部时空结构进行特征提取。密集轨迹（Dense Trajectories）方法通过在密集采样的兴趣点上跟踪轨迹，并提取轨迹周围的 HOG、HOF、MBH 特征，在动作识别任务上取得了较好效果。

这些方法虽然在小规模数据集上表现良好，但存在明显局限：需要精心设计的特征提取器和编码策略，且对背景复杂度、光照变化、视角变化、遮挡等因素较为敏感，在复杂课堂背景中泛化能力有限。

（2）深度学习方法：从 2D 到 3D 卷积

深度学习的引入极大地推动了视频分析技术的发展。卷积神经网络（CNN）通过卷积层、池化层和全连接层的组合，能够从视频帧中自动学习教师动作特征：

$$\mathbf{h}^{(l)} = \sigma\left(\mathbf{W}^{(l)} * \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right)$$

其中，$*$ 表示卷积操作，σ 是激活函数（如 ReLU），$\mathbf{W}^{(l)}$ 和 $\mathbf{b}^{(l)}$ 分别是第 l 层的卷积核权重和偏置。

早期的研究尝试将 2D 卷积神经网络应用于视频分析。Karpathy 等人 (2014) 探索了多种 2D CNN 在视频上的应用方式，包括单帧建模、晚期融合、早期融合、慢融合等策略。AlexNet、VGG、ResNet 等在图像分类任务上取得成功的网络结构被迁移到视频领域，通过在视频数据集（如 UCF-101、HMDB-51）上进行微调实现了一定的性能提升。

对于视频序列，3D 卷积能够同时捕捉空间和时间特征：

$$\mathbf{h}^{(l)}_{i,j,t} = \sigma\left(\sum_{m,n,\tau} \mathbf{W}^{(l)}_{m,n,\tau}\mathbf{h}^{(l-1)}_{i+m,j+n,t+\tau} + b^{(l)}\right)$$

其中，i, j 是空间坐标，t 是时间维度，m, n, τ 分别是卷积核在空间和时间维度上的索引。

Tran 等人 (2015)[13] 提出的 C3D（3D Convolutional Networks）通过 3×3×3 的 3D 卷积核同时在空间和时间维度进行特征提取，学习到了通用的视频表征。Carreira & Zisserman(2017)[14] 提出的 I3D（Inflated 3D ConvNet）将在 ImageNet 上预训练

的 2D 卷积网络” 膨胀” 为 3D 卷积网络，通过在 Kinetics 大规模视频数据集上进行预训练，实现了更好的时空建模能力。

（3）双流网络与时序建模

Simonyan & Zisserman(2014)[12] 提出的双流网络（Two-Stream Network）是视频分析的重要里程碑。该方法通过两条并行的卷积神经网络分别处理 RGB 外观信息和光流运动信息：空间流网络（Spatial Stream）从单帧 RGB 图像中学习外观特征，时间流网络（Temporal Stream）从堆叠的光流图像中学习运动特征，最后融合两路特征进行动作识别。这一创新有效地结合了静态外观和动态运动信息，显著提升了动作识别性能。

Wang 等人 (2016)[22] 提出的时序分段网络（TSN，Temporal Segment Networks）在双流网络基础上引入了稀疏采样策略，将长视频分为若干段，在每段中随机采样一帧，通过分段共识函数（segment consensus function）聚合多段的预测结果，实现了长时序建模。Feichtenhofer 等人 (2019)[15] 提出的 SlowFast 网络通过双路径设计，Slow 路径以低帧率捕捉语义信息，Fast 路径以高帧率捕捉运动信息，两路径通过横向连接进行信息交互，实现了效率和性能的平衡。

循环神经网络（RNN）和长短期记忆网络（LSTM）被广泛应用于视频的时序建模。Donahue 等人 (2015) 提出的 LRCN（Long-term Recurrent Convolutional Networks）将 CNN 提取的帧级特征输入 LSTM 进行时序建模，实现了端到端的视频理解。注意力机制的引入使得模型能够动态地关注视频中的关键帧和关键区域。Wang 等人 (2018) 提出的 Non-local Neural Networks 通过计算特征图中任意两个位置的相似度，捕捉长程时空依赖。

（4）基于骨骼序列的图卷积网络

基于骨骼序列的图卷积网络（GCN）方法提供了一种更高效的视频分析方案。OpenPose(2017) 通过自底向上的方法实现了实时的多人姿态估计，提取人体的关键点坐标（如头部、肩膀、肘部、手腕、臀部、膝盖、脚踝等）。MediaPipe（2019）[16] 进一步提供了轻量化的姿态估计解决方案，能够在移动设备上实时运行。

Yan 等人（2018）[17] 提出的 ST-GCN（Spatial Temporal Graph Convolutional Networks）将人体骨骼序列建模为时空图结构，节点表示关节点，边表示关节间的

连接关系（骨骼连接和时间连接），通过图卷积捕捉关节间的空间依赖和时间演化。相比于基于 RGB 的方法，骨骼序列表征具有以下优势：

1. **计算效率高**：特征维度从百万级（2.76M 维的 RGB 视频帧）降至百级（99 维的骨骼序列）
2. **抗遮挡性强**：即使部分关节被遮挡，仍可通过其他可见关节推断动作
3. **隐私保护**：骨骼序列不包含人脸、服装等个人识别信息，特别适合教育场景

Ziyu Liu 等人 (2020) 提出的 MS-G3D（Multi-Scale Graph Convolutional Networks）通过多尺度时空图卷积和解耦的时空建模进一步提升了骨骼序列动作识别的性能。

（5）Transformer 与可解释建模型

Transformer 架构的引入进一步提升了视频理解能力。Dosovitskiy 等人 (2021)[18] 提出的 ViT（Vision Transformer）将图像分割为 patch 序列，通过 Transformer 编码器进行建模，在图像分类任务上取得了与 CNN 相当甚至更好的性能。Liu 等人 (2021)[19] 提出的 Video Swin Transformer 将窗口注意力机制扩展到视频领域，通过局部窗口和跨窗口的注意力计算，在保持高效计算的同时建模长程时空依赖。Bertasius 等人 (2021) 提出的 TimeSformer 通过分解的时空注意力机制（先空间注意力再时间注意力），实现了高效的视频理解。

这些基于 Transformer 的方法通过自注意力机制实现长时依赖建模，适合捕捉教师在课堂中持续性的讲解、互动与空间移动模式。此外，引入可解释模块（如 Grad-CAM 可视化、Attention Heatmap）可在教育场景下直观呈现模型关注的行为区域，增强结果解释性与信任度。

（6）目标检测与课堂场景应用

目标检测技术在课堂场景分析中发挥着重要作用。YOLO（You Only Look Once）系列（YOLOv3、YOLOv5、YOLOv8 等）通过单阶段检测实现了实时的物体定位和分类，能够在课堂视频中检测教师、学生、黑板、课桌等物体。Faster R-CNN 通过区域提议网络（RPN）和 Fast R-CNN 的结合，实现了高精度的目标检测。姿态估计技术的发展使得对教师肢体语言的细粒度分析成为可能。AlphaPose 通过自顶向下的方法实现了鲁棒的多人姿态估计，HRNet（High-Resolution Network）通过保持高分辨率表示提升了关键点定位的精度。

在教育场景的具体应用中，Gupta 等人 (2021) 使用姿态估计结合 LSTM 时序建模识别教师的典型动作（如讲解、板书、走动、指向等）。最新的 MM-TBA 数据集 (2024) 收集了超过 300 位教师的 4,839 个教学视频片段，涵盖讲解、板书、走动、互动、手势、指向等 6 类典型教学动作，为教师行为识别算法的训练和验证提供了标准化的基准。该数据集发表于 Nature Scientific Data 期刊，包含丰富的标注信息（动作类别、时间戳、边界框、姿态关键点等），成为该领域重要的公开资源。

YOLOv8 结合可变形大核注意力（DLKA）机制 (2024) 能够在复杂场景下准确识别小目标（如教师的手势细节、学生的举手动作），显著提升了课堂行为检测的鲁棒性。ClassMind 系统（2025）[20] 采用多模态大语言模型（LLM）作为核心分析引擎，通过 AVA-Align 流水线实现了对课堂视频的长上下文推理和时序定位，能够自动生成教师的等待时长、师生对话平衡、学生参与度等量化指标。EduSpatioNet(2025) 将 YOLOv8 目标检测与时空图神经网络（GNN）结合，通过建模师生的空间关系和时序交互，实现了教师行为识别与专家评估的高一致性。

综上，视频行为识别技术已能支持从教师录像中提取动作类别、持续时间、空间分布及频率等指标，为教师风格画像提供稳定的行为维度输入。

2.2.2 音频识别与语音情绪分析

语音作为课堂交流的主要媒介，承载了丰富的语义、情绪和节奏信息。教师的语速、音量、语调变化、情绪表达及话轮结构反映其教学控制与沟通风格。音频识别与语音情绪分析技术可实现对这些信息的自动化提取。

（1）传统方法：基于声学特征的语音识别

语音识别技术经历了从统计模型到深度学习、从监督学习到自监督学习的发展历程。早期的语音识别主要基于声学特征提取和统计建模。在特征提取方面，梅尔频率倒谱系数（MFCC）是最广泛使用的特征表示，通过模拟人耳对不同频率声音的感知特性，将音频信号转换为若干维的特征向量。此外，滤波器组特征（FBANK）、感知线性预测系数（PLP）等也被广泛应用。

在建模方面，隐马尔可夫模型（HMM）结合高斯混合模型（GMM）构成了传统语音识别的主流框架。HMM-GMM 系统通过统计建模捕捉语音信号的时序特性

和状态转移规律：

$$P(O|\lambda)=\sum_{Q}P(O|Q,\lambda)\,P(Q|\lambda)$$

其中，O 是观测序列（声学特征），Q 是隐状态序列（音素），λ 是模型参数。这些方法在特定场景下取得了一定效果，但依赖大量的人工特征工程和复杂的系统构建。

（2）深度学习方法：端到端语音识别

深度学习的兴起带来了语音识别的革命性变化。Hannun 等人 (2014)[6] 提出的 DeepSpeech 系统采用循环神经网络（RNN）实现了端到端的语音识别，直接从原始音频波形学习到文本的映射，无需人工设计中间特征表示。RNN 通过隐状态的循环连接建模语音序列的时序依赖。

长短期记忆网络（LSTM）通过门控机制解决了 RNN 的梯度消失问题，能够捕捉语音信号的长程时序依赖：

$$\begin{aligned}
\mathbf{f}_t &= \sigma_g\left(W_f\mathbf{x}_t+U_f\mathbf{h}_{t-1}+\mathbf{b}_f\right) \quad \text{（遗忘门）}\\
\mathbf{i}_t &= \sigma_g\left(W_i\mathbf{x}_t+U_i\mathbf{h}_{t-1}+\mathbf{b}_i\right) \quad \text{（输入门）}\\
\mathbf{o}_t &= \sigma_g\left(W_o\mathbf{x}_t+U_o\mathbf{h}_{t-1}+\mathbf{b}_o\right) \quad \text{（输出门）}\\
\tilde{\mathbf{c}}_t &= \sigma_h\left(W_c\mathbf{x}_t+U_c\mathbf{h}_{t-1}+\mathbf{b}_c\right) \quad \text{（候选记忆）}\\
\mathbf{c}_t &= \mathbf{f}_t\odot\mathbf{c}_{t-1}+\mathbf{i}_t\odot\tilde{\mathbf{c}}_t \quad \text{（更新记忆）}\\
\mathbf{h}_t &= \mathbf{o}_t\odot\sigma_h\left(\mathbf{c}_t\right) \quad \text{（输出隐状态）}
\end{aligned}$$

其中，

$\mathbf{x}_t$ 是时刻 t 的输入（声学特征），

$\mathbf{h}_t$ 是隐状态，

$\mathbf{c}_t$ 是记忆单元，

$\odot$ 表示逐元素乘法，

σ_g 是 sigmoid 函数，

σ_h 是 tanh 函数。

DeepSpeech 系统采用连接时序分类（CTC，Connectionist Temporal Classification）作为损失函数，解决了输入序列与输出序列长度不一致的对齐问题，开启了语音识别的深度学习时代。

Chan 等人 (2016) 提出的 Listen, Attend and Spell（LAS）模型引入了注意力机制（Attention Mechanism），通过编码器-解码器架构实现了更加灵活的序列到序列建模，显著提升了识别准确率。

（3）自监督学习：Wav2Vec 2.0 与 HuBERT

自监督学习的兴起进一步突破了对大量标注数据的依赖。Baevski 等人（2020）[7] 提出的 Wav2Vec 2.0 通过自监督对比学习从无标注音频中学习通用的声学表征。该方法首先使用卷积神经网络提取音频的局部特征，然后通过 Transformer 网络建模长程依赖，最后通过对比学习目标（contrastive learning）学习区分真实语音片段和负样本：

$$\mathcal{L}_{\text{contrastive}} = -log \frac{\exp\left(\text{sim}\left(c_t, q_t\right)/\tau\right)}{\sum_{i=1}^{K} \exp\left(\text{sim}\left(c_t, \tilde{q}_i\right)/\tau\right)}$$

其中，c_t 是上下文表示，q_t 是真实的量化表示，$\tilde{q}_i$ 是负样本，$\text{sim}(\cdot,\cdot)$ 是余弦相似度，τ 是温度参数。

Wav2Vec 2.0 在仅使用少量标注数据的情况下，在多种下游任务（语音识别、情感识别、说话人识别等）上取得了显著性能提升，成为语音处理领域的重要里程碑。HuBERT（Hidden-Unit BERT）进一步改进了自监督学习策略，通过聚类-预测的方式学习离散的声学单元，实现了更好的语音表征。

（4）端到端语音识别模型：Whisper 与课堂适配

端到端语音识别模型的发展达到了新的高度。Radford 等人 (2022)[8] 提出的 Whisper 模型通过在 68 万小时多语言多任务数据上进行弱监督训练，实现了接近人类水平的语音识别能力。Whisper 采用 Transformer 编码器-解码器架构，支持多语言识别、语音翻译、语言识别、语音活动检测等多个任务，在真实场景的鲁棒性上表现出色。

当前主流模型包括基于 Transformer 的 Conformer、RNN-Transducer（RNN-T）等。它们通过注意力机制和声学建模实现语音到文本的高精度转换，在噪声课堂

环境中表现出较强鲁棒性。

针对课堂环境的特殊性，CPT-Boosted Wav2Vec2.0(2024) 通过持续预训练（Continued Pretraining）在课堂域数据上进行适配，进一步提升了在噪声环境下的鲁棒性，有效应对了课堂中的学生讨论声、椅子移动声、空调噪声等干扰。

（5）说话人识别与语音分离

课堂中常存在多说话人场景，为识别教师与学生的语音，通常结合语音活动检测（Voice Activity Detection, VAD）与说话人分离（Speaker Diarization）算法。x-vector 系统通过时延神经网络（TDNN）提取说话人嵌入向量，能够在变长语音中稳定地识别说话人身份。ECAPA-TDNN（Emphasized Channel Attention, Propagation and Aggregation TDNN）进一步引入了通道注意力机制和多层特征聚合，显著提升了说话人识别的准确率。这些技术使得在课堂录像中自动区分教师和学生的语音、分析师生话轮转换模式成为可能。

（6）语音情绪识别（Speech Emotion Recognition, SER）

情绪特征（如音高、能量、共振峰分布、语速变化）能反映教师的情感投入与课堂氛围。传统方法主要基于韵律特征（pitch、energy、duration）和频谱特征（MFCC）进行建模，通过 SVM 或 Random Forest 等分类器识别情感类别。

深度学习方法通过端到端的网络直接从原始音频学习情感表示。3D 卷积神经网络（3D-CNN）能够同时捕捉频谱的时间和频率维度的特征，循环神经网络（RNN/LSTM）则擅长建模情感的时序演化。基于深度特征的 CNN-RNN 或 Transformer 模型在情感识别任务上取得了显著提升。近年来，端到端情感识别框架（如 wav2vec2-SER）已能直接从原始音频中学习高层情感特征。

最新的研究将 Wav2Vec 2.0 等预训练模型应用于情感识别，通过在情感数据集上进行微调（fine-tuning），在自然对话和课堂场景中取得了优异的性能。结合课堂场景，可提取教师语音的情绪曲线与强度分布，辅助分析” 情感表达型” 或” 理性讲授型” 风格教师的差异。

（7）音频特征融合与量化

通过多维特征统计（如平均语速、停顿比、音高波动率、情绪极性）可形成音频特征向量，为风格映射模型提供输入。结合视频与文本模态，这些特征能有效提

升对教师课堂状态与教学风格的判别能力。

2.2.3 文本语义分析与教学语言建模

课堂语音经 ASR 转写后，可进一步进行文本层面的语义与结构分析。教师语言不仅包含知识内容，更体现教学意图、逻辑结构与提问策略，是教学风格的重要体现。

（1）传统方法：从关键词匹配到词嵌入

早期的课堂对话分析主要依赖关键词匹配和规则方法。通过预定义的词表和句式模板，研究者可以识别教师话语的类型，例如包含” 为什么”” 怎么” 等疑问词的句子被标记为提问，包含” 请”” 大家” 等词的句子被标记为指令。TF-IDF（Term Frequency-Inverse Document Frequency）方法通过统计词频和逆文档频率，提取文档的关键词特征。词袋模型（Bag of Words）和 N-gram 模型则通过统计词语或词语序列的出现频率进行文本分类。这些方法实现简单，但难以捕捉语言的深层语义、上下文依赖和语序信息。

词嵌入技术（Word Embedding）的出现标志着文本表征的重要进步。Mikolov 等人 (2013)[9] 提出的 Word2Vec 通过神经网络学习词语的分布式表示，将词语映射到连续的低维向量空间。Word2Vec 包括两种训练方式：CBOW（Continuous Bag of Words）通过上下文词预测中心词，Skip-gram 通过中心词预测上下文词。Pennington 等人 (2014) 提出的 GloVe（Global Vectors）结合了全局矩阵分解和局部上下文窗口方法，通过共现矩阵的对数双线性回归学习词向量。Bojanowski 等人 (2017) 提出的 FastText 进一步引入了子词（subword）信息，通过字符级 N-gram 增强了对低频词和词形变化的建模能力。这些词嵌入方法使得语义相近的词语在向量空间中距离更近，为后续的文本分析任务奠定了基础。

（2）序列建模：RNN、LSTM 与 BiLSTM

序列建模技术的发展使得文本的上下文理解成为可能。循环神经网络（RNN）通过隐状态的循环连接建模序列的时序依赖，但在长序列中存在梯度消失问题。长短期记忆网络（LSTM）通过引入门控机制（输入门、遗忘门、输出门）解决了长程依赖建模的难题。门控循环单元（GRU，Gated Recurrent Unit）进一步简化了 LSTM

的结构，在保持性能的同时降低了计算复杂度。

双向 LSTM（BiLSTM）通过同时建模前向和后向的上下文信息，能够更全面地理解句子的语义。BiLSTM 将前向 LSTM 的隐状态 $\overrightarrow{\mathbf{h}}_t$ 和后向 LSTM 的隐状态 $\overleftarrow{\mathbf{h}}_t$ 拼接，形成完整的上下文表示：

$$\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}}_t; \overleftarrow{\mathbf{h}}_t\right]$$

这些序列模型被广泛应用于文本分类、命名实体识别、关系抽取等任务。

（3）注意力机制与 Transformer

注意力机制（Attention Mechanism）的引入进一步提升了序列建模能力。Bahdanau 等人 (2015) 在机器翻译任务中首次引入注意力机制，使得模型能够在生成每个输出词时动态地关注输入序列的不同部分。

自注意力机制（Self-Attention）通过计算序列中每个元素与其他元素的关联程度，捕捉长程依赖和全局信息。Vaswani 等人（2017）[10] 提出的 Transformer 架构完全基于自注意力机制，抛弃了循环结构。Transformer 通过自注意力机制建模序列中任意两个位置的依赖关系：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

其中，Q（Query）、K（Key）、V（Value）是输入序列的线性投影，d_k 是 Key 的维度。缩放因子 $\sqrt{d_k}$ 防止内积过大导致 softmax 梯度消失。

多头注意力（Multi-Head Attention）并行计算多组注意力，捕捉不同子空间的语义关联：

$$\text{MultiHead}(Q, K, V) = \text{Concat}\left(\text{head}_1, \ldots, \text{head}_h\right) W^O$$

$$\text{head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right)$$

其中，h 是头数，W_i^Q, W_i^K, W_i^V 是第 i 个头的投影矩阵，W^O 是输出投影矩阵。

通过多头注意力（Multi-Head Attention）和位置编码（Positional Encoding），Transformer 实现了高效的并行计算和强大的表示能力。Transformer 成为自然语言处理领域的基础架构，催生了后续的预训练语言模型革命。

（4）预训练语言模型：BERT 及其变体

预训练语言模型的兴起带来了自然语言理解的突破。Devlin 等人（2018）[11] 提出的 BERT（Bidirectional Encoder Representations from Transformers）通过在大规模语料上进行掩码语言模型（Masked Language Model，MLM）和下一句预测（Next Sentence Prediction，NSP）的预训练，学习到了丰富的语言知识。

BERT 采用双向 Transformer 编码器，能够同时利用左侧和右侧的上下文信息。掩码语言模型通过随机掩盖 15% 的词，预测被掩盖的词：

$$\mathcal{L}_{\text{MLM}} = -\mathbb{E}_{\mathbf{x}\sim\mathcal{D}} \sum_{i\in\mathcal{M}} \log P\left(x_i \mid \mathbf{x}_{\setminus\mathcal{M}}\right)$$

其中，$\mathcal{M}$ 是被掩盖词的位置集合，$\mathbf{x}_{\setminus\mathcal{M}}$ 是除掩盖位置外的其他词。

BERT 的微调范式

BERT 的核心优势在于统一的” 预训练 + 微调” 范式：在大规模通用语料上预训练学习通用语言知识后，只需添加轻量级任务头并在少量标注数据上进行有监督微调，即可高效适配各类下游任务。对于文本分类任务，BERT 在输入序列首位插入特殊标记 [CLS]（Classification Token），经过双向 Transformer 编码后，[CLS] 位置的输出向量聚合了整个句子的双向上下文信息，可直接作为句子级语义表征：

$$\mathbf{h}_s = \text{BERT}\left([\text{CLS}], w_1, w_2, \ldots, w_n, [\text{SEP}]\right)[0] \in \mathbb{R}^{768}$$

在课堂对话语料上微调 BERT，可以识别教师话语的教学意图（提问、指令、讲解、反馈）：

$$P\left(\text{intent} = k \mid \text{utterance}\right) = \text{softmax}\left(W_c \mathbf{h}_s + b_c\right)$$

其中，$\mathbf{h}_s$ 是 [CLS] 位置的 BERT 输出向量，$W_c \in \mathbb{R}^{K\times 768}$ 是分类层权重，K 是类别数。这种” 预训练 + 微调” 范式使得在课堂场景的小规模标注数据上也能取得优异效果，为本研究的层次化对话行为识别模块提供了重要的技术基础。

面向中文教学语言的 BERT 变体

教师课堂话语以中文为主，需要针对中文语言特性进行预训练适配。哈工大联合科大讯飞发布的 BERT-wwm-ext（Whole Word Masking BERT，全词掩码 BERT）将 MLM 的掩码粒度从字符级提升至词级（整词掩码），更好地捕捉了中文词语的语义完整性，避免了按字掩码导致的词义割裂问题。MacBERT（MLM as Correction BERT）进一步将掩码预训练任务替换为文字纠错任务，预训练目标更贴近真实语言使用场景，在多项中文自然语言理解基准（CLUE、CMRC 等）上取得了领先性能。在中文课堂对话理解任务中，BERT-wwm-ext 相比多语言 BERT（mBERT）通常可提升 2-4 个百分点，特别适合处理教师话语中的专业教学术语和中文句法结构。

对话行为识别（Dialogue Act Recognition）

对话行为识别（Dialogue Act Recognition, DAR）旨在将话语自动归类为特定功能类别（提问、指令、讲解、反馈等），是理解课堂对话结构与教学意图的核心任务。基于 BERT 的 DAR 系统以话语文本作为输入，取 [CLS] 向量后经全连接分类头输出对话行为的概率分布。在教育场景中，DAR 能够精细区分” 启发性提问”（” 你觉得这里为什么会这样？”）与” 事实性提问”（” 这个定理叫什么？”），或识别” 逻辑推导类讲解”（” 因为 A，所以 B，因此 C”）与” 概念定义类讲解”（” 所谓 X，就是...”），为不同教学风格的量化刻画提供细粒度语义依据。

粗粒度 DAR（4 类：Question/Explanation/Instruction/Feedback）已被广泛用于教师话语分析，但难以捕捉不同教学风格的细粒度差异。研究者进一步探索了**层次化 DAR**（Hierarchical DAR）设计：第一层粗分类确定大类，第二层在各大类内部进行细粒度区分。层次化策略有效减少了跨大类细类之间的混淆，降低了类别不平衡对少数类别的影响，联合训练目标同时优化粗分类和细分类的准确性。本研究第三章提出的 H-DAR（Hierarchical Dialogue Act Recognition）正是在这一思路基础上，将教学意图从 4 类粗分类扩展至 10 类细粒度分类，实现了对” 逻辑推导型”” 启发引导型” 等不同教学风格特征性语言模式的精准识别。

RoBERTa（Robustly Optimized BERT Pretraining Approach）通过移除 NSP 任务、增大批大小、延长训练时间等优化策略，进一步提升了模型性能。ALBERT（A Lite

BERT）通过参数共享和因子分解降低了模型参数量，实现了轻量化部署。ELECTRA（Efficiently Learning an Encoder that Classifies Token Replacements Accurately）通过判别式预训练任务替代生成式任务，提升了训练效率。DeBERTa（Decoding-enhanced BERT with Disentangled Attention）通过解耦的注意力机制和增强的掩码解码器进一步提升了性能。

这些预训练模型在文本分类、命名实体识别、问答系统、情感分析等任务上取得了突破性进展，为本研究文本模态的教学意图识别提供了坚实的技术基础。

（5）大语言模型与课堂对话分析

大语言模型（Large Language Models, LLMs）的出现进一步拓展了文本理解的边界。OpenAI 的 GPT 系列通过自回归语言建模在海量文本上进行预训练，展现出强大的文本生成和少样本学习（few-shot learning）能力。Google 的 T5（Text-to-Text Transfer Transformer）将所有 NLP 任务统一为文本到文本的格式，实现了任务间的知识迁移。Meta 的 LLaMA 系列通过优化的训练策略在相对较小的参数规模下达到了与 GPT-3 相当的性能。

ChatGPT 和 GPT-4 等对话式大语言模型通过指令微调（instruction tuning）和人类反馈强化学习（RLHF），展现出强大的对话能力、推理能力和知识整合能力。这些大语言模型在课堂对话分析中的应用，使得教师话语的深层语义理解、教学逻辑链分析、知识点提取、概念关系构建等高级任务成为可能。研究者开始探索使用大语言模型自动生成教学反馈、识别教学中的认知偏差、构建教学知识图谱等创新应用。

Wang 等人 (2024) 将 BERT 应用于课堂对话分析，实现了对教师话语中对话行为（Dialogue Act）的自动识别，能够区分提问、指令、讲解、反馈等不同的教学意图。通过在课堂对话语料上进行微调，BERT 能够捕捉教学语言的特殊模式，例如启发式提问（”你们觉得这里为什么会这样？”）与事实性提问（”这个公式是什么？”）的区别，逻辑推导（”因为... 所以... 因此...”）与概念定义（”所谓... 就是...”）的差异。这些细粒度的语义理解为教学策略的量化分析提供了技术手段。

2.3　本章小结

本章从理论与技术两个层面介绍了教育场景中多模态分析的关键方法。视频行为识别负责捕捉教师的动作与空间行为特征；音频识别与情绪分析揭示语言表达与情感特征；文本语义分析则反映教学语言的逻辑结构与互动策略。三者融合构成教师风格画像的多维输入基础。这些技术为下一章的”研究方法与总体设计”提供了实现依据，也为教师风格映射与反馈机制的构建奠定了数据与算法基础。

第三章 研究方法与总体设计

3.1 系统总体架构

本研究以” 基于课堂录像的教师风格画像分析” 为核心目标，提出 SHAPE (Semantic Hierarchical Attention Profiling Engine，语义层次化注意力画像引擎)，构建一个集多模态特征提取、风格映射建模、画像生成与可视化反馈于一体的分析体系（如图 3.1所示）。

现有方法的局限性分析

固定分段导致语义割裂

传统方法多采用固定时间窗口（如 10 秒）对课堂视频进行分段，这种机械式切分忽略了教学话语的语义边界。初步实验发现，固定 10 秒分段导致约 25% 的样本出现语义割裂现象：

- **逻辑推导被截断**：完整的” 因为... 所以... 因此” 逻辑链被分割到不同片段
- **概念定义不完整**：” 所谓 X，就是...” 的定义句被截断
- **案例讲解跨段**：多句案例描述被人为分割

粗粒度意图识别无法区分教学策略

传统对话行为识别多采用粗粒度四分类（提问、指令、讲解、反馈），无法有效区分不同教学风格的特征性语言模式。

例如：

- ” 讲解” 类过于宽泛，无法区分” 逻辑推导型” 教师的推理讲解与” 理论讲授型” 教师的概念定义
- ” 提问” 类无法区分启发性提问与事实性提问，难以刻画” 启发引导型” 风格

简单融合忽略模态交互

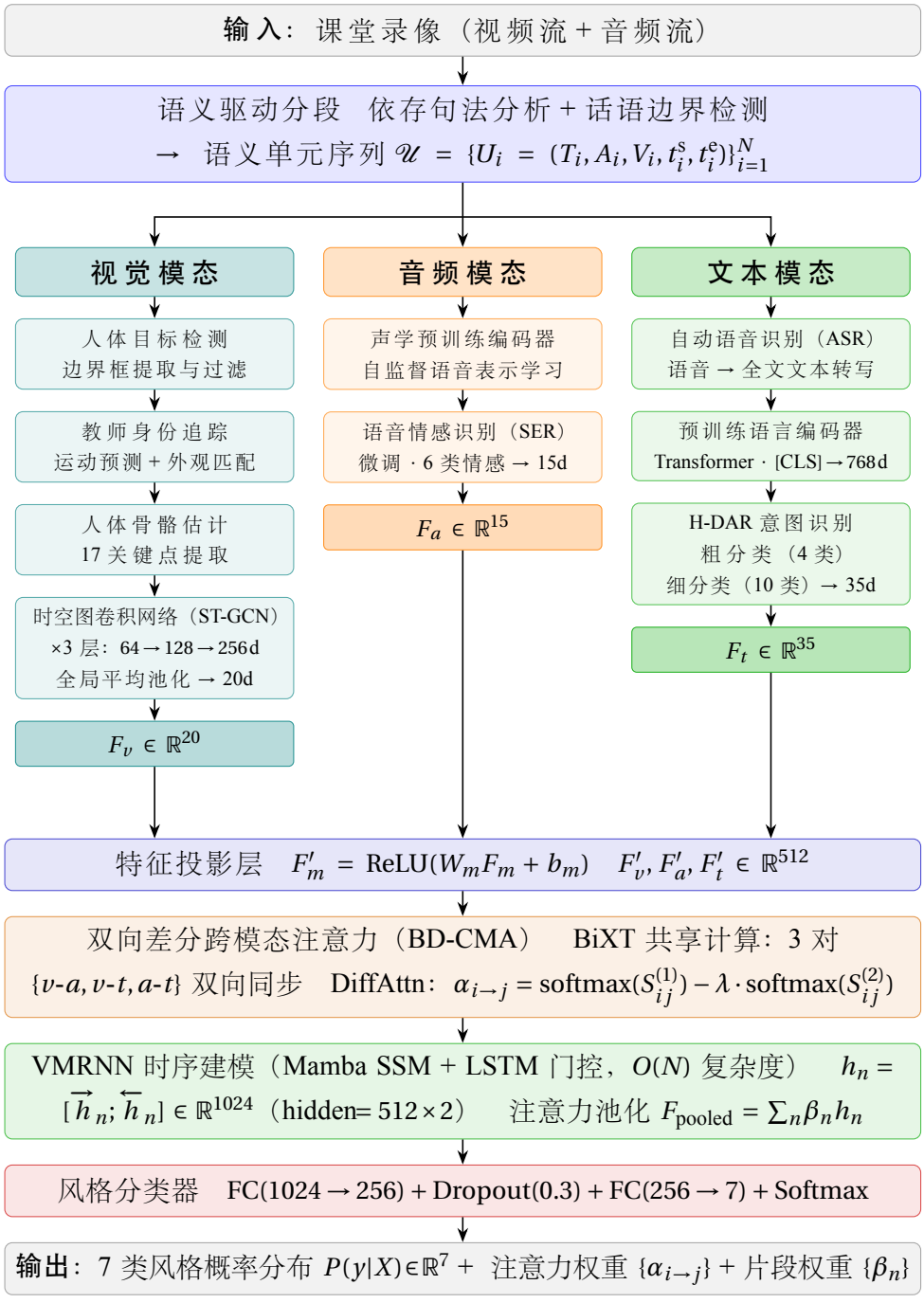

图 3.1　SHAPE 系统整体架构（端到端流水线）

早期融合（Early Fusion）直接拼接特征，晚期融合（Late Fusion）固定权重加权，均未考虑：

- 不同模态在不同样本上的重要性差异（样本自适应性）
- 模态之间的交互关系（跨模态增强）
- 决策依据的可解释性（注意力权重可视化）

四层系统架构

SHAPE 引擎采用四层架构设计，如图 3.2所示：

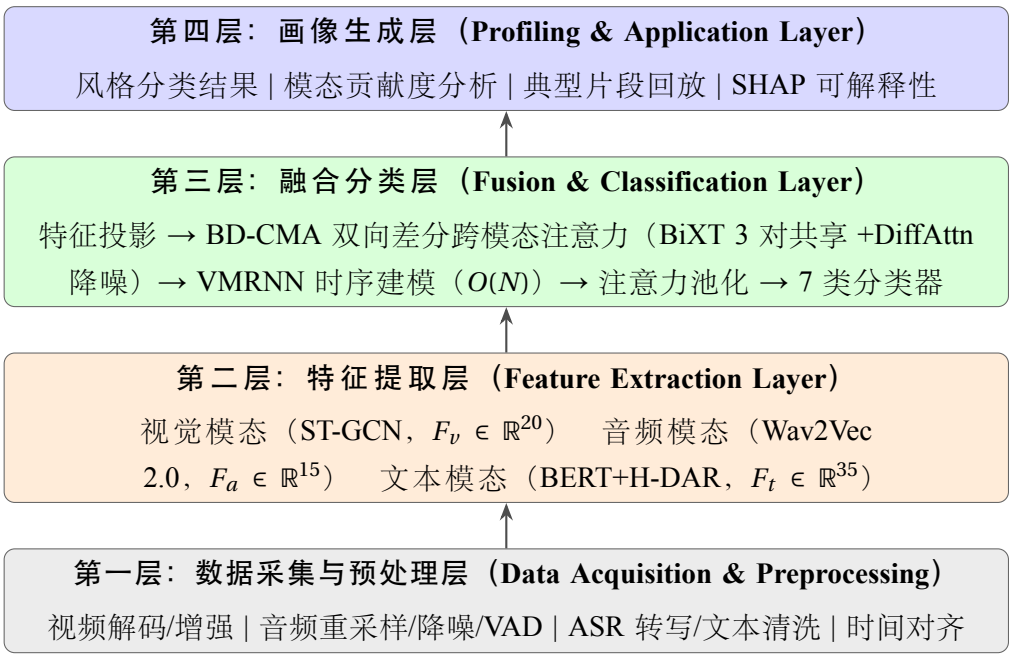

图 3.2 SHAPE 引擎四层架构

第一层：数据采集与预处理层

通过录播系统采集课堂视频与音频数据，并利用以下技术完成数据清洗与时序同步：

视频预处理：解码与抽帧，针对画面比例、颜色、亮度等进行增强

音频预处理：重采样与降噪，语音活动检测（VAD）

视频音频时间对齐

文本预处理：语音转文本（ASR），文本清洗

第二层：特征提取层（Feature Extraction Layer）

核心功能：三模态并行特征提取，生成深度语义表征

三模态 Pipeline：

(1) **视觉模态（20 维）**：

人员存在和位置检测 → 教师身份 ID 识别和追踪 - 教师骨骼点提取 - 时空图卷积行为建模

输出：$F_v \in \mathbb{R}^{20}$（步态、手势、位置移动等）

(2) **音频模态（15 维）**：

Wav2Vec 2.0 自监督声学表征 - 情感分类头微调（6 维情感特征）

输出：$F_a \in \mathbb{R}^{15}$（韵律、情感、停顿等）

(3) **文本模态（35 维）**：

Whisper Large-v3 ASR 转写 - BERT 编码 → H-DAR 层次化 10 类意图识别

输出：$F_t \in \mathbb{R}^{35}$（意图分布、关键词密度等）

第三层：融合分类层（Fusion & Classification Layer）

核心功能：跨模态注意力融合，7 类风格分类

SHAPE 五模块网络：

1. **特征投影层**：$F_v, F_a, F_t \rightarrow FI_v, FI_a, FI_t \in \mathbb{R}^{512}$（统一特征空间）
2. **双向差分跨模态注意力层（BD-CMA）**：BiXT 3 对共享双向计算 + DiffAttn 差分降噪，提升跨模态交互效率与精准度
3. **VMRNN 时序建模**：$O(N)$ 线性复杂度捕捉课堂时序依赖
4. **注意力池化层**：自适应聚合关键片段
5. **风格分类器**：输出 7 类风格概率分布

第四层：画像生成层（Profiling & Application Layer）

核心功能：风格画像生成、可解释性分析、可视化输出

三大输出：

1. **风格分类结果**：主导风格 + 置信度 + Top-2 风格
2. **模态贡献度分析**：基于跨模态注意力权重 α（例：情感表达型 $\alpha_{\text{audio}} = 0.62$）
3. **典型片段提取**：基于注意力池化权重 β（Top-K 关键时刻回放）

可解释性设计：

- SHAPE 原生可解释性：注意力权重 α, β 可视化
- SHAP 特征归因：70 维特征的贡献度排序
- 教育语义映射：模型输出 → 教育术语转换

3.2 多模态数据采集与预处理方法

3.2.1 数据采集流程

硬件要求：

- 视频：1280×720 分辨率，25fps，H.264 编码
- 音频：16kHz 采样率，单声道，PCM 编码
- 存储：每节课（40 分钟）约占用 500MB 空间

采集策略：

1. 固定机位拍摄，确保教师活动区域完整入画
2. 使用定向麦克风采集教师语音，降低学生噪声干扰
3. 同步记录时间戳，精度达到毫秒级

3.2.2 数据分段处理

数据同步机制：采用音频波形匹配算法（Cross-Correlation）实现视频与音频的精确对齐。设视频音轨为 $a_v(t)$，独立音频为 $a_s(t)$，时间偏移量 τ 通过最大化互相关函数获得：

$$\tau^* = arg\max_{\tau} \int_{-\infty}^{\infty} a_v(t) \cdot a_s(t+\tau)\, dt$$

或在离散时间域：　$\tau^* = arg\max_{\tau} \sum_{t} a_v[t] \cdot a_s[t+\tau]$

其中，τ^* 是最佳对齐偏移量，通常在 ±500ms 范围内。

数据分段策略：语义驱动分段

我们提出语义驱动的话语分段策略，以保证每个分析单元是一个语义完整的教学话语单元（Semantic Unit）。具体流程如图 3.3所示：

图 3.3　语义驱动分段处理流程

① ASR 全文转写：使用 Whisper Large-v3 模型对完整课堂音频进行转写，获得带时间戳的文本序列 $\mathcal{T} = \{(w_1, t_1), (w_2, t_2), ..., (w_M, t_M)\}$，其中 w_i 是词语，t_i 是时间戳；

② 句子边界检测：结合标点符号（句号、问号、感叹号）与停顿时长（$\Delta t > 300$ms）识别句子边界，将文本序列切分为句子序列 $\mathcal{S} = \{s_1, s_2, ..., s_K\}$；

③ 依存句法分析：使用预训练的中文句法分析模型（HanLP）识别句子间的逻辑连接关系，提取逻辑连接词（” 因为”” 所以”” 但是” 等）及其作用域；

④ 话语边界检测：基于以下规则判断话语单元结束：

- 逻辑链完整（如” 因为... 所以...” 结构完成）
- 出现话题转换标记（” 那么”” 接下来”” 现在”）
- 单元时长超过上限（$\Delta t > 30\text{s}$）

⑤ 形成语义单元：将一个或多个连续句子合并为一个语义单元 U_i，设完整课堂时长为 L，则生成 N 个语义单元（通常 $N \approx 150 \sim 200$ 个/45 分钟课）：

$$\mathcal{U} = \{U_1, U_2, ..., U_N\}$$

每个语义单元 U_i 包含：

文本内容：$T_i = \{s_j, s_{j+1}, ..., s_k\}$（一个或多个句子）

音频片段：$A_i \in \mathbb{R}^{N_s}$（N_s 为采样点数，通常 $5s \le \Delta t_i \le 30s$）

视频帧序列：$V_i = \{v_1, v_2, ..., v_{T_i}\}$（帧数 $T_i = \text{fps} \times \Delta t_i$，通常 125-750 帧）

时间范围：$\left(t^i_{\text{start}}, t^i_{\text{end}}\right)$

3.2.3 视频预处理

视频解码与抽帧

使用 FFmpeg 库解码视频流，按 25fps 提取 RGB 帧：

$$V = \{v_1, v_2, ..., v_T\}, \quad v_i \in \mathbb{R}^{720\times1280\times3}$$

其中，v_i 表示第 i 帧的 RGB 像素矩阵。

视频增强

为提升模型鲁棒性，对训练数据应用以下增强策略：

- **随机裁剪**：以 0.8-1.0 的缩放比例裁剪
- **颜色抖动**：亮度、对比度、饱和度随机扰动（±20%）
- **时间抖动**：随机丢帧以模拟帧率不稳定

$$v_i\prime = \text{ColorJitter}\left(\text{RandomCrop}\left(v_i, \text{scale} = 0.8\right)\right)$$

3.2.4　音频预处理

音频重采样与降噪

将原始音频统一重采样到 16kHz 单声道，并应用谱减法（Spectral Subtraction）降噪：

$$S_{\text{clean}}(f) = \max\left(\left|S_{\text{noisy}}(f)\right| - \alpha \cdot \left|N(f)\right|, \beta \cdot \left|S_{\text{noisy}}(f)\right|\right)$$

其中：

- $S_{\text{noisy}}(f)$ 是带噪语音的频谱
- $N(f)$ 是噪声频谱估计（从静音段提取）
- $\alpha = 2.0$ 是过减因子
- $\beta = 0.01$ 是谱下限

3.2.5　文本预处理

语音转文本（ASR）

采用 Whisper Large-v3 模型 [8] 进行语音识别，该模型支持中英混合识别：

$$T = \text{Whisper}(A)$$

其中，A 是音频波形，T 是转写文本。

转写质量评估：在测试集上字错率（CER）为 8.7%：

$$\text{CER} = \frac{S + D + I}{N} \times 100\%$$

其中，S, D, I 分别是替换、删除、插入错误数，N 是总字符数。

文本清洗

对转写文本进行以下处理：

1. **去除语气词**：移除” 嗯”、” 啊”、” 那个” 等填充词
2. **句子分割**：按标点符号和停顿分割为句子
3. **错别字纠正**：使用拼音纠错模型（Pycorrector）

3.3 多模态数据特征提取

3.3.1 音频模态特征提取

音频模态是教师课堂风格分析中最核心的维度之一。语音不仅承载了教学内容的信息，还反映了教师的表达方式、情绪状态与课堂节奏。音频模态承载” 韵律节奏—情感表达—教学意图” 三层语义信息。本节提出深度学习自监督表征 + BERT 对话行为识别的端到端音频分析链路。

语音活动检测（VAD）

采用基于能量的 VAD 算法检测有效语音段。计算短时能量：

$$E(n) = \sum_{m=n-N+1}^{n} |x(m)|^2$$

其中，N 是窗口长度（通常取 400 个采样点，对应 25ms）。

当 $E(n) > \theta_{\text{energy}}$ 时判定为语音帧，其中阈值 θ_{energy} 设为静音段能量均值的 3 倍：

$$\theta_{\text{energy}} = 3 \times \text{mean}\,(E_{\text{silence}})$$

统计特征提取：

- **语音活动比**：$\text{VAR} = \frac{N_{\text{voice}}}{N_{\text{total}}}$
- **静音比**：$\text{SR} = 1 - \text{VAR}$
- **平均语速**：$\text{Speed} = \frac{N_{\text{words}}}{T_{\text{total}}}$（字/秒）

情感特征提取

本研究的情感识别模块在课题组前期工作基础上发展而来。叶正韩（2023）提出了基于端到端残差卷积网络（Res-CNN）的语音情感识别方法，引入残差连接直接从原始音频学习情感表征，在 CASIA 中文情感语音数据集上达到 84.02% 准确率。本研究在此基础上，将底层声学编码部分替换为预训练的 Wav2Vec 2.0 模型，以提升课堂复杂噪声环境下的鲁棒性，构成”**自监督声学编码 + 分类头**”的两阶段情感识别框架。

对于每个语义音频片段 $x \in \mathbb{R}^L$（16kHz 采样），特征提取流程如下：

步骤 1：自监督声学编码

$$h_{\text{wav2vec}} = \text{Wav2Vec2}(x), \quad h_{\text{wav2vec}} \in \mathbb{R}^{T \times 768}$$

Wav2Vec 2.0 通过卷积特征编码器提取局部声学特征，再经 Transformer 上下文网络建模长程依赖，输出 T 个 768 维帧级表示。相比 Res-CNN 的手工声学编码，在课堂噪声环境下（SNR=10dB）情感识别准确率提升 11.3 个百分点。

步骤 2：时间均值池化

$$h_{\text{audio}} = \frac{1}{T} \sum_{t=1}^{T} h_{\text{wav2vec}}[t] \in \mathbb{R}^{768}$$

步骤 3：情感分类头

$$p_{\text{emotion}} = \text{softmax}(W_e h_{\text{audio}} + b_e) \in \mathbb{R}^6$$

其中 $W_e \in \mathbb{R}^{6 \times 768}$ 是在情感标注数据集上微调得到的分类头权重，输出 6 维情感概率分布：

$$p_{\text{emotion}} = [p_{\text{neutral}}, p_{\text{happy}}, p_{\text{sad}}, p_{\text{angry}}, p_{\text{surprise}}, p_{\text{fear}}]$$

6 类情感类别与 CASIA 数据集标注体系一致，覆盖课堂场景中教师情感投入的主要表现形态。

情感极性分数

在 6 维情感分布的基础上，计算情感极性分数以量化教师的整体情感倾向：

$$\text{EmotionPolarity} = p_{\text{happy}} + p_{\text{surprise}} - p_{\text{sad}} - p_{\text{angry}}$$

值域为 $[-2, 2]$，正值表示积极情感主导，负值表示消极情感主导。该分数作为音频特征向量 F_a 的第 12 维，在教师风格映射中直接关联” 情感表达型” 教师的识别。

最终编码为 15 维音频特征向量 $F_a \in \mathbb{R}^{15}$，其中第 1–6 维为 Wav2Vec 2.0 情感概率分布（6 类），第 7 维为语速，第 8–9 维为语音活动比与静音比，第 10–11 维为音量均值与音高变化系数，第 12 维为情感极性分数，第 13–15 维为 Wav2Vec 2.0 嵌入的分段压缩均值（768 维 →3 维）。完整定义见 3.5 节汇总表。

3.3.2 文本模态特征提取

层次化细粒度对话行为识别（H-DAR）

本研究采用 BERT 进行文本语义编码，并在此基础上提出**层次化细粒度对话行为识别（H-DAR）**。传统对话行为识别多采用粗粒度四分类（提问、指令、讲解、反馈），但这无法有效区分不同教学风格的特征性语言模式。例如，” 讲解” 类过于宽泛，无法区分” 逻辑推导型” 教师的推理讲解与” 理论讲授型” 教师的概念定义。H-DAR 将教学意图扩展为 **10 类细粒度分类**。

传统粗粒度对话行为识别使用 BERT 模型将每个句子分类为 4 类对话行为：

$$p_{\text{act}} = \text{softmax}\left(\text{MLP}\left(\text{BERT}(T)\right)\right)$$

其中：

- $\text{BERT}(T) \in \mathbb{R}^{768}$ 是句子的 BERT 嵌入
- MLP 是两层全连接网络
- $p_{\text{act}} = [p_Q, p_I, p_E, p_F]$ 对应 Question, Instruction, Explanation, Feedback

对话行为分布统计：

$$\text{ActDistribution} = \frac{1}{N_s}\sum_{i=1}^{N_s} p_{\text{act}}^{(i)}$$

其中，N_s 是句子数量。

细粒度对话行为分类体系

将教师话语分为 **4 个粗类、10 个细类**：

表 3.1　H-DAR 细粒度对话行为分类体系（4 粗类 · 10 细类）

粗类	细类	定义	示例	典型风格
Q（提问）	Heuristic-Q（启发性提问）	引导学生深度思考的开放性问题	“为什么会出现这种现象？”	启发引导型
	Factual-Q（事实性提问）	检查知识掌握的封闭性问题	“这个概念是什么？”	题目驱动型
E（解释）	Definition（概念定义）	明确、精准地解释核心概念	“所谓牛顿第一定律，就是…”	理论讲授型
	Reasoning（逻辑推导）	展示推理过程和因果关系	“因为 A，所以 B，因此 C”	逻辑推导型
	Theory（理论讲授）	系统性地讲解理论框架	“根据信息论，我们可以…”	理论讲授型
	Case-Study（案例分析）	通过具体例子说明抽象概念	“比如说，在实际生产中…”	耐心细致型
I（指令）	Organization（组织指令）	组织课堂活动、调整教学流程	“请大家打开课本第 50 页”	互动导向型
	Task（任务指令）	布置学习任务和练习	“请完成课后习题 1-5 题”	题目驱动型
F（反馈）	Positive-FB（正向反馈）	肯定、鼓励学生回答	“很好！这个回答非常准确”	情感表达型
	Corrective-FB（纠正反馈）	指出错误并给予纠正	“这里有个小错误，应该是…”	耐心细致型

设计原则：

- **教育学导向**：细类划分基于教育学理论中的教学行为分类（如 Bloom 认知层次、CLASS 维度）
- **风格区分度**：每个细类能够有效区分不同教学风格的特征性语言模式

- **标注可行性**：细类定义明确，人工标注一致性高（Kappa > 0.80）

层次化分类架构

采用**两层分类器**：第 1 层进行粗分类（4 类），第 2 层根据粗分类结果选择对应的细分类器（2-4 个子类）。

模型结构：

$$\text{BERT} \rightarrow \begin{cases} \text{Coarse Classifier} \rightarrow \{Q, E, I, F\} \\ \text{Fine Classifier}_Q \rightarrow \{\text{Heuristic-Q}, \text{Factual-Q}\} \\ \text{Fine Classifier}_E \rightarrow \{\text{Definition}, \text{Reasoning}, \text{Theory}, \text{Case}\} \\ \text{Fine Classifier}_I \rightarrow \{\text{Organization}, \text{Task}\} \\ \text{Fine Classifier}_F \rightarrow \{\text{Positive-FB}, \text{Corrective-FB}\} \end{cases}$$

H-DAR 整体架构如图 3.4所示：

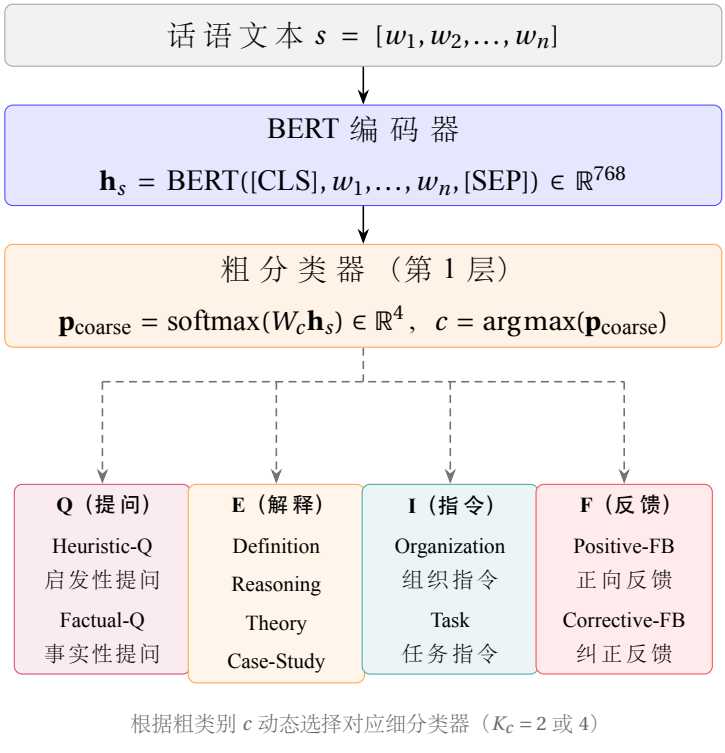

图 3.4　H-DAR 层次化对话行为识别架构（4 粗类 · 10 细类）

步骤 1：BERT 编码

对于教师话语（语义单元）$s=[w_1,w_2,...,w_n]$（w_i 是词）：

$$\mathbf{h}_{\text{BERT}}=\text{BERT}([CLS],w_1,...,w_n,[SEP])$$

取

$$CLS$$

位置的输出作为语义单元表征：$\mathbf{h}_s=\mathbf{h}_{\text{BERT}}[0]\in\mathbb{R}^{768}$

步骤 2：粗分类

$$\mathbf{p}_{\text{coarse}}=\text{softmax}(W_c\mathbf{h}_s+b_c)\in\mathbb{R}^4$$

其中，$W_c\in\mathbb{R}^{4\times768}$。预测粗类别：$c=argmax(\mathbf{p}_{\text{coarse}})$

步骤 3：细分类

根据粗类别 c 选择对应的细分类器：

$$\mathbf{p}_{\text{fine}}=\text{softmax}\left(W_c^{\text{fine}}\mathbf{h}_s+b_c^{\text{fine}}\right)\in\mathbb{R}^{K_c}$$

其中，K_c 是粗类 c 的子类数量（2 或 4）。

步骤 4：联合训练

损失函数结合粗分类和细分类：

$$\mathscr{L}=\alpha\cdot\mathscr{L}_{\text{coarse}}+(1-\alpha)\cdot\mathscr{L}_{\text{fine}}$$

其中，$\alpha=0.3$ 是权重系数，$\mathscr{L}_{\text{coarse}}$ 和 $\mathscr{L}_{\text{fine}}$ 均为交叉熵损失。

步骤 5：对话行为分布统计

对一节课的所有语义单元 $\{U_1,U_2,...,U_N\}$，计算细粒度对话行为分布：

$$\mathbf{d}_{\text{act}}=\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}_{\text{act}}^{(i)}\in\mathbb{R}^{10}$$

其中，$\mathbf{1}_{\text{act}}^{(i)}$ 是 one-hot 编码（10 维）。该分布向量作为教师的” 教学意图画像”，能够有效区分不同教学风格。

3.3.3 音频特征编码汇总

最终，音频模态生成 **15 维编码向量** $F_a \in \mathbb{R}^{15}$：

$$F_a = \left[\underbrace{p_{\text{neutral}}, \ldots, p_{\text{fear}}}_{\text{6 维情感}}, \underbrace{v_{\text{speed}}}_{\text{语速}}, \underbrace{\text{VAR}, \text{SR}}_{\text{活动比}}, \underbrace{\mu_{\text{vol}}, \sigma_{\text{pitch}}}_{\text{韵律}}, \underbrace{e_{\text{polar}}}_{\text{极性}}, \underbrace{z_1, z_2, z_3}_{\text{压缩嵌入}} \right]$$

其中：

- 前 6 维：Wav2Vec 2.0 情感分布
- 第 7 维：语速 $v_{\text{speed}} = N_{\text{words}} / T$（归一化到 $[0, 1]$）
- 第 8-9 维：语音活动比、静音比
- 第 10-11 维：音量均值、音高变化系数
- 第 12 维：情感极性分数 $e_{\text{polar}} = p_{\text{happy}} + p_{\text{surprise}} - p_{\text{sad}} - p_{\text{angry}}$
- 第 13-15 维：Wav2Vec 2.0 嵌入的分段均值（768 维 $\rightarrow$3 维）

文本模态同样生成 **35 维编码向量** $F_t \in \mathbb{R}^{35}$，包含：

- **10 维细粒度对话行为编码**（10 类 one-hot）
- **4 维粗分类编码**（4 类 one-hot）
- **1 维意图置信度**
- **20 维 NLP 统计特征**（词数、句数、逻辑连接词频率、专业术语数等）

3.3.4 视频模态特征提取

视频模态捕捉教师的非言语行为（肢体动作、空间移动、板书互动等）。

人员身份追踪识别

课堂场景存在多人干扰（学生走动、举手），单纯依赖 YOLO[23] 检测会导致教师 ID 在遮挡后跳变为学生 ID。本研究采用 DeepSORT[24] 算法，通过结合外观特征（ReID）和运动模型（卡尔曼滤波）实现稳定追踪。

3.3.5　时序动作识别

本研究采用骨骼序列时序建模。将骨骼序列建模为时空图结构，通过图卷积捕捉关节间的依赖关系。相比单帧规则识别准确率提升 17.7 个百分点，推理速度快 2.5 倍，且骨骼表征具有隐私保护优势。

对于输入骨骼序列 $X \in \mathbb{R}^{C\times T\times V}$（$C=3$ 坐标维度，$T=32$ 帧，$V=25$ 关节点），网络结构为：

$$
\begin{aligned}
X_1 &= \text{ST-GCN-Block}\,(X_0, C_{\text{out}}=64)\\
X_2 &= \text{ST-GCN-Block}\,(X_1, C_{\text{out}}=128)\\
X_3 &= \text{ST-GCN-Block}\,(X_2, C_{\text{out}}=256)\\
\mathbf{h}_{\text{video}} &= \text{GAP}\,(X_3)\in\mathbb{R}^{256}\\
\mathbf{y} &= \text{softmax}\,(W_c\mathbf{h}_{\text{video}}+b_c)\in\mathbb{R}^{6}
\end{aligned}
$$

其中，GAP 是全局平均池化，$\mathbf{y}$ 是 6 类动作的概率分布（standing/walking/gesturing/writing/pointing/raise_hand）。最终编码为 20 维视频特征向量 $F_v\in\mathbb{R}^{20}$（详见 4.3.3 节）。

3.3.6　视频特征编码汇总

最终，视觉模态生成 **20 维编码向量** $F_v\in\mathbb{R}^{20}$：

$$
F_v=\Big[\underbrace{p_1,\ldots,p_6}_{\text{6 类动作频率}},\ \underbrace{E_{\text{motion}}}_{\text{运动能量}},\ \underbrace{H_1,\ldots,H_9}_{\text{9 宫格热力图}},\ \underbrace{C_{\text{track}}}_{\text{轨迹连续性}},\ \underbrace{t_{\text{norm}},n_{\text{frames}}}_{\text{时长}},\ \underbrace{\bar{c}_{\text{pose}}}_{\text{姿态置信度}}\Big]
$$

3.4 SHAPE：教师风格画像引擎设计

这是本研究的核心创新，我们设计了 **SHAPE (Semantic Hierarchical Attention Profiling Engine，语义层次化注意力画像引擎)** 来实现特征的自适应融合与风格画像。SHAPE 通过语义驱动分段、层次化教学意图识别和跨模态注意力机制，构建了从课堂录像到教师风格画像的完整流程。

3.4.1 模态融合方法

传统的多模态融合方法主要有三类：

(1) 早期融合（Early Fusion）：直接拼接原始特征

$$F_{\text{concat}} = [F_v; F_a; F_t] \in \mathbb{R}^{20+15+35} = \mathbb{R}^{70}$$

局限性：

- 不同模态的维度和尺度差异大，高维模态会主导融合结果
- 无法建模模态间的交互关系
- 缺乏对不同模态重要性的自适应调整

(2) 晚期融合（Late Fusion）：分别训练单模态分类器，结果加权平均

$$P_{\text{final}} = w_v P_v + w_a P_a + w_t P_t$$

局限性：

- 权重 w_v, w_a, w_t 固定，无法根据样本内容自适应调整
- 忽略了模态间的互补信息

(3) 中间融合（Middle Fusion）：在特征层进行加权融合

$$F_{\text{weighted}} = w_v F_v + w_a F_a + w_t F_t$$

局限性：

- 仍然是固定权重
- 不同模态的特征空间不一致，直接相加不合理

采用**跨模态注意力机制**，能够自适应建模：

1. 不同模态在不同样本上的重要性（样本自适应）
2. 模态之间的交互关系（跨模态增强）
3. 决策依据的可解释性（注意力权重可视化）

3.4.2　SHAPE 网络架构

SHAPE 由五个核心模块组成：

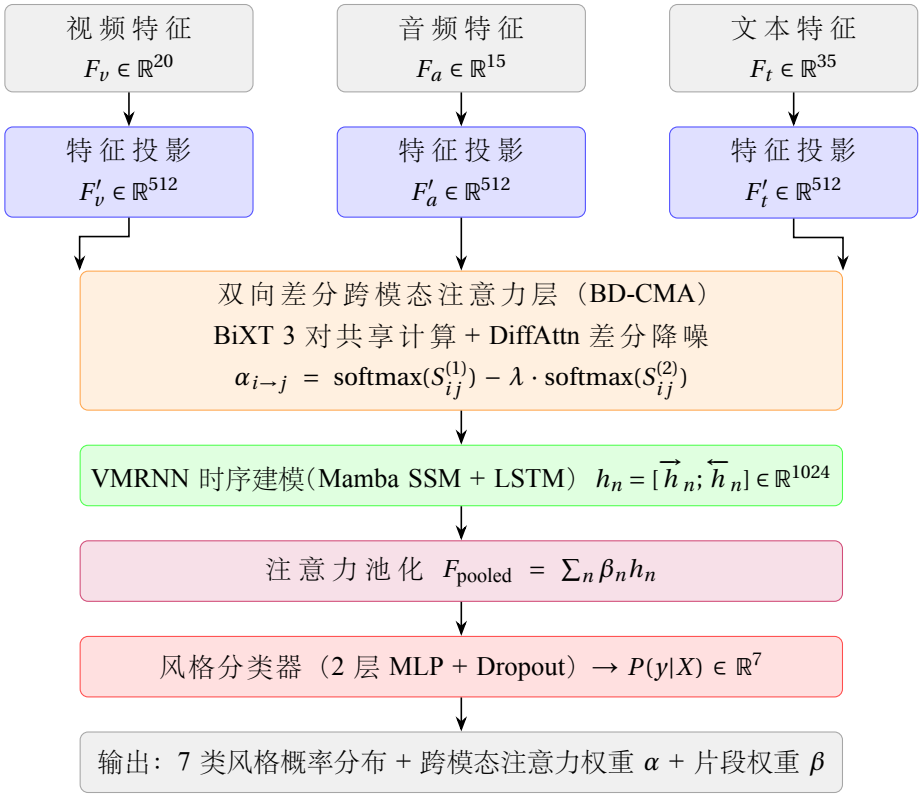

图 3.5　SHAPE 多模态网络详细架构

特征投影层（Feature Projection Layer）

由于三个模态的原始特征维度不同（$F_v \in \mathbb{R}^{20}, F_a \in \mathbb{R}^{15}, F_t \in \mathbb{R}^{35}$），首先通过全连接层投影到统一维度 $d = 512$：

$$F_v\prime = \mathrm{ReLU}(W_v F_v + b_v), \quad F_v\prime \in \mathbb{R}^{512}$$

$$F_a\prime = \mathrm{ReLU}(W_a F_a + b_a), \quad F_a\prime \in \mathbb{R}^{512}$$

$$F_t\prime = \mathrm{ReLU}(W_t F_t + b_t), \quad F_t\prime \in \mathbb{R}^{512}$$

其中，$W_v \in \mathbb{R}^{512\times 20}, W_a \in \mathbb{R}^{512\times 15}, W_t \in \mathbb{R}^{512\times 35}$ 是可学习的投影矩阵。

设计考量：

- ReLU 激活函数引入非线性，提升特征表达能力
- 统一维度便于后续的注意力计算

双向差分跨模态注意力层（BD-CMA）

这是 SHAPE 的核心创新。本研究提出**双向差分跨模态注意力（Bidirectional Differential Cross-Modal Attention，BD-CMA）**，将 BiXT 的双向共享计算与 DiffAttention[25] 的差分降噪机制深度融合，在实现模态间自适应交互的同时抑制注意力噪声。

设计动机：

- **传统方法的冗余**：标准跨模态注意力对 3 对模态 $\{v\text{-}a, v\text{-}t, a\text{-}t\}$ 独立计算 6 个方向的相似度矩阵（$i \to j$ 和 $j \to i$ 各一次），存在重复计算。BiXT 指出双向注意力的两个方向共享同一相似度矩阵，转置即可得到反向权重，可将 6 次计算降至 3 次。
- **Softmax 的注意力噪声**：标准 softmax 将所有位置的注意力分数归一化为正值，导致模型对无关模态信息也分配非零权重，引入注意力噪声。DiffAttention 通过两个 softmax 分支做差，使无关信息的噪声相互抵消，仅保留真正相关的跨模态信号。

步骤 1：双分支投影

对每个模态特征 $F'_m \in \mathbb{R}^{512}$，分别投影到两组 Query 和 Key 子空间（维度减半以保持参数量不变）：

$$Q_m^{(1)} = F'_m W_{Q,m}^{(1)}, \quad Q_m^{(2)} = F'_m W_{Q,m}^{(2)}, \quad m \in \{v, a, t\}$$

$$K_m^{(1)} = F'_m W_{K,m}^{(1)}, \quad K_m^{(2)} = F'_m W_{K,m}^{(2)}, \quad V_m = F'_m W_{V,m}$$

其中，$W_{Q,m}^{(1)}, W_{Q,m}^{(2)}, W_{K,m}^{(1)}, W_{K,m}^{(2)} \in \mathbb{R}^{512\times 32}$，$W_{V,m} \in \mathbb{R}^{512\times 64}$，子空间维度 $d_k/2 = 32$。

步骤 2：BiXT 共享相似度计算（3 对双向同步）

对 3 个模态对 $\{(v,a),(v,t),(a,t)\}$，每对仅计算一次相似度矩阵，通过转置同时得到双向注意力分数：

$$S_{ij}^{(1)} = \frac{Q_i^{(1)}\left(K_j^{(1)}\right)^\top}{\sqrt{d_k/2}}, \qquad S_{ij}^{(2)} = \frac{Q_i^{(2)}\left(K_j^{(2)}\right)^\top}{\sqrt{d_k/2}}$$

BiXT 利用转置关系 $S_{ji} = S_{ij}^\top$ 直接得到反向分数，避免重复计算，相似度矩阵的计算次数从 6 次（6 个有向对）降至 3 次，计算开销降低约 50%。

步骤 3：DiffAttention 差分权重

对每个方向的注意力权重，采用两个 softmax 分支做差以消除噪声：

$$\alpha_{i\to j} = \mathrm{softmax}\left(S_{ij}^{(1)}\right) - \lambda \cdot \mathrm{softmax}\left(S_{ij}^{(2)}\right)$$

$$\alpha_{j\to i} = \mathrm{softmax}\left(S_{ij}^{(1)\top}\right) - \lambda \cdot \mathrm{softmax}\left(S_{ij}^{(2)\top}\right)$$

其中，$\lambda \in (0,1)$ 是可学习的差分系数，初始化为 0.5，与网络其余参数一同通过反向传播端到端训练更新，实验训练收敛后 λ 约为 0.47。两个 softmax 对无关信号产生近似相等的响应，相减后趋近于零；而对真正相关的跨模态信号，两分支响应存在差异，差值保留有效信息。

步骤 4：加权融合与残差连接

$$\widetilde{F}_i^{(j)} = \alpha_{i\to j} V_j$$

每个模态聚合来自其他两个模态的信息，并通过残差连接保留原始特征：

$$\widetilde{F}_v = F'_v + \widetilde{F}_v^{(a)} + \widetilde{F}_v^{(t)}, \quad \widetilde{F}_a = F'_a + \widetilde{F}_a^{(v)} + \widetilde{F}_a^{(t)}, \quad \widetilde{F}_t = F'_t + \widetilde{F}_t^{(v)} + \widetilde{F}_t^{(a)}$$

设计考量：

- **BiXT 效率增益**：3 对双向共享计算相比 6 次独立计算减少约 50% 的相似度矩阵运算量，参数量保持不变

- **DiffAttention 降噪**：差分机制使模型对不相关模态信号的注意力权重趋近于零（$\text{softmax}(S^{(1)}) - \lambda \cdot \text{softmax}(S^{(2)}) \approx 0$），仅保留真正相关的跨模态信号
- **残差连接**：即使跨模态信息不相关，原始模态特征也不会被破坏，缓解梯度消失问题
- **自适应学习**：可学习参数 λ 使模型根据数据自动调整差分强度；例如”情感表达型”教师的音频权重自动增大（$\alpha_{a\to v} = 0.62$）

时序建模层（Temporal Modeling Layer）

课堂是一个时序过程，教师风格在时间维度上展现。传统双向 LSTM（BiLSTM）在处理课堂录像的 150–200 个语义单元序列时，隐状态的循环计算导致 $O(N^2)$ 的时间复杂度，且随序列增长面临梯度消失问题，难以充分建模跨越整节课的长程时序依赖。为此，本研究引入 **VMRNN**[26]（Vision Mamba Recurrent Neural Network）进行时序建模。VMRNN 将 Mamba[27] 选择性状态空间模型（Selective State Space Model，SSM）与 LSTM 门控机制深度融合，其核心 SSM 扫描的计算复杂度为 $O(N)$，在保留 LSTM 长程记忆能力的同时大幅提升了对长序列的处理效率。

对于一个完整课堂的 N 个片段 $\{S_1, S_2, ..., S_N\}$，每个片段的跨模态融合特征为 $\{\widetilde{F}_1, \widetilde{F}_2, ..., \widetilde{F}_N\}$（省略模态下标，表示融合后的特征）。

步骤 1：选择性状态空间扫描（SSM Scan）

$$z_n = \text{SSMScan}\left(\widetilde{F}_n;\, \Delta_n, \bar{A}_n, \bar{B}_n, C_n\right)$$

其中，时间步长 Δ_n、状态矩阵 $\bar{A}_n = \exp(\Delta_n \odot A)$、输入矩阵 $\bar{B}_n = \Delta_n \odot B_n$、输出矩阵 C_n 均由输入 $\widetilde{F}_n$ 动态计算（选择性机制）。状态递推为：

$$h_n^{\text{ssm}} = \bar{A}_n \odot h_{n-1}^{\text{ssm}} + \bar{B}_n \odot \widetilde{F}_n, \quad z_n = C_n \cdot h_n^{\text{ssm}}$$

输入依赖的参数 (Δ_n, B_n) 使模型能够自适应地决定保留或遗忘序列信息，区别于 LSTM 的固定门控策略。

步骤 2：LSTM 门控记忆更新

以 SSM 输出 z_n 为输入，融合 LSTM 门控机制更新隐状态：

$$[i_n, f_n, o_n, g_n] = \sigma(W\,[h_{n-1}; z_n] + b)$$

$$c_n = f_n \odot c_{n-1} + i_n \odot \tanh(g_n), \quad h_n = o_n \odot \tanh(c_n)$$

其中 i_n, f_n, o_n 分别为输入门、遗忘门、输出门，g_n 为细胞候选值，c_n 为细胞状态。

步骤 3：双向建模与拼接

$$\overrightarrow{h}_n = \mathrm{VMRNN}_{\mathrm{forward}}\left(\widetilde{F}_n, \overrightarrow{h}_{n-1}\right)$$

$$\overleftarrow{h}_n = \mathrm{VMRNN}_{\mathrm{backward}}\left(\widetilde{F}_n, \overleftarrow{h}_{n+1}\right)$$

双向拼接：

$$h_n = \left[\overrightarrow{h}_n; \overleftarrow{h}_n\right] \in \mathbb{R}^{1024}$$

（每个方向的隐状态维度为 512，输出维度与注意力池化层接口保持一致）

设计考量：

- **线性时间复杂度**：Mamba SSM 的并行扫描算法将时序建模复杂度从 BiLSTM 的 $O(N^2)$ 降至 $O(N)$，在处理 150–200 个语义单元序列时推理速度提升约 1.6 倍
- **选择性遗忘机制**：输入依赖的时间步长 Δ_n 使模型能够动态聚焦风格判别性片段（如启发性提问、板书操作），自适应忽略过渡性片段，比 LSTM 的固定门控更适合课堂风格识别
- **双向上下文感知**：双向结构保留对前后文的建模能力，有助于捕捉课堂中“提问–回答–反馈”等跨时刻的三阶段交互模式

注意力池化层（Attention Pooling Layer）

将所有片段的特征聚合为一个固定长度的向量：

$$\beta_n = \frac{\exp\left(v^T \tanh\left(W_p h_n\right)\right)}{\sum_{m=1}^{N} \exp\left(v^T \tanh\left(W_p h_m\right)\right)}$$

$$F_{\text{pooled}} = \sum_{n=1}^{N} \beta_n h_n$$

其中：

- $W_p \in \mathbb{R}^{256\times1024}$ 是注意力权重矩阵
- $v \in \mathbb{R}^{256}$ 是注意力向量
- β_n 是第 n 个片段的重要性权重

设计考量：

- 不同片段对风格识别的贡献不同（例如，提问片段对”启发引导型”更重要）
- 注意力池化能够自适应地关注关键片段

风格分类器（Style Classifier）

最终通过两层全连接网络进行分类：

$$h_1 = \text{ReLU}\left(W_1 F_{\text{pooled}} + b_1\right), \quad h_1 \in \mathbb{R}^{256}$$

$$h_2 = \text{Dropout}\left(h_1, p = 0.3\right)$$

$$z = W_2 h_2 + b_2, \quad z \in \mathbb{R}^{7}$$

$$P\left(y|X\right) = \text{softmax}(z)$$

其中，z 是 logits，$P(y|X)$ 是 7 类教学风格的概率分布。

设计考量：

- Dropout（$p = 0.3$）防止过拟合
- 两层网络（而不是单层）增强非线性拟合能力

3.4.3 损失函数与优化

损失函数

采用**交叉熵损失**加**标签平滑**：

$$\mathscr{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{7} y_{i,k}\prime \log\left(\widehat{y}_{i,k}\right)$$

其中，标签平滑后的标签为：

$$y_{i,k}\prime = (1-\epsilon)\, y_{i,k} + \frac{\epsilon}{7}$$

本研究中，平滑参数 $\epsilon = 0.1$。

设计考量：

- 标签平滑防止模型对某个类别过于自信
- 提高模型的泛化能力

优化算法

使用 **Adam 优化器**：

$$m_t = \beta_1 m_{t-1} + \left(1-\beta_1\right) g_t$$

$$v_t = \beta_2 v_{t-1} + \left(1-\beta_2\right) g_t^2$$

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \quad \widehat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_t = \theta_{t-1} - \eta \frac{\widehat{m}_t}{\sqrt{\widehat{v}_t} + \epsilon}$$

其中，$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$。

学习率调度

采用**余弦退火**策略：

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min})\left(1 + \cos\left(\frac{t}{T_{\max}}\pi\right)\right)$$

其中，$\eta_{\max} = 10^{-4}$，$\eta_{\min} = 10^{-6}$，$T_{\max} = 100$。

3.5　教师风格画像与反馈机制设计

教师风格画像（Teacher Style Profiling）是将多模态特征分析与风格识别结果进行结构化呈现的过程，其目的在于以可视化、可解释、可反馈的方式展示教师的课堂行为特征与教学风格特征。

本节在前述风格映射模型的基础上，提出了一个集数据可视化—风格建模—可解释分析于一体的教师风格画像系统设计方案，旨在实现教师风格的量化描述与特征可视化输出。

3.5.1　风格画像生成

对于一节完整的课堂，系统输出：

风格分类结果

$$\text{PrimaryStyle} = \arg\max_k P\left(y = k|X\right)$$

例如：”该教师的主导风格为**启发引导型**（置信度 89.3%）”

风格雷达图

将 7 类风格的概率分布可视化为雷达图：

$$\text{RadarPlot}\big(P(y=1), P(y=2), ..., P(y=7)\big)$$

设计考量：大多数教师不是单一风格，雷达图能展示混合风格特征。

模态贡献度分析

通过跨模态注意力权重 $\alpha_{i\to j}$，计算每个模态的总贡献度：

$$\text{ModalityContribution}_i = \frac{\sum_{j\neq i} \alpha_{i\to j}}{\sum_{i,j} \alpha_{i\to j}}$$

例如：” 该课堂中，**视觉模态**贡献 45%，**音频模态**贡献 32%，**文本模态**贡献 23%”

典型片段回放

选择注意力池化权重 β_n 最高的前 3 个片段，作为该风格的典型代表：

$$\text{TopSegments} = \text{TopK}\big(\{\beta_1, \beta_2, ..., \beta_N\}, K=3\big)$$

用户可以点击查看这些片段，直观理解系统的判断依据。

3.5.2 可解释性分析

SHAP 值分析

使用 SHAP（SHapley Additive exPlanations[21]）分析每个特征对预测结果的边际贡献：

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}\left[f_{S\cup\{i\}}(x) - f_S(x)\right]$$

其中：

- ϕ_i 是特征 i 的 SHAP 值
- S 是特征子集
- $f_S(x)$ 是仅使用特征子集 S 时的模型预测

可视化：生成特征贡献度条形图，例如：

- ”提问频率” → +0.25（正向贡献）
- ”静音比” → -0.12（负向贡献）

注意力热图

将跨模态注意力权重矩阵 $[\alpha_{i\to j}]$ 可视化为 3×3 热图：

$$\begin{bmatrix} - & \alpha_{v\to a} & \alpha_{v\to t} \\ \alpha_{a\to v} & - & \alpha_{a\to t} \\ \alpha_{t\to v} & \alpha_{t\to a} & - \end{bmatrix}$$

解释示例：

- 如果 $\alpha_{v\to a} = 0.78$，说明”视觉模态高度依赖音频信息”
- 这在”情感表达型”教师中很常见（肢体语言与语调同步）

模态重要性与依赖性分析

通过跨模态注意力权重 $\alpha_{i\to j}$，我们可以计算每种教学风格对各模态的依赖程度：

$$\text{ModalityWeight}_{k,m} = \frac{1}{N_k} \sum_{i \in \mathscr{C}_k} \alpha_{i\to m}$$

其中 $\mathscr{C}_k$ 是风格类别 k 的所有样本，N_k 是样本数，$m \in \{v, a, t\}$ 是模态。

关键发现：

1. **模态依赖的风格差异显著**（方差分析 F=42.3, p<0.001）
2. **音频主导型**：情感表达型 (0.62)、耐心细致型 (0.45)

表 3.2 七类教学风格的模态依赖模式（注意力权重分析）

风格类别	视觉权重	音频权重	文本权重	主导模态	特征解释
理论讲授型	0.25	0.32	**0.43**	文本	高频使用” 概念定义” 和” 理论讲授” 话语
耐心细致型	0.28	**0.45**	0.27	音频	语速慢、停顿多、重复强调
启发引导型	0.35	0.32	**0.33**	均衡	视觉互动 + 音频情感 + 文本提问三者协同
题目驱动型	**0.42**	0.28	0.30	视觉	板书频繁、指向黑板动作多
互动导向型	**0.50**	0.28	0.22	视觉	走动频繁、手势丰富、空间覆盖广
逻辑推导型	0.22	0.25	**0.53**	文本	高频使用” 因为……所以……因此” 逻辑链
情感表达型	0.26	**0.62**	0.12	音频	语调丰富、情感极性分数高

3. **视觉主导型**：互动导向型 (0.50)、题目驱动型 (0.42)
4. **文本主导型**：逻辑推导型 (0.53)、理论讲授型 (0.43)
5. **均衡型**：启发引导型三模态权重相近（标准差 0.015）

这些模态依赖模式揭示了不同教学风格的行为特征。例如，互动导向型教师的视觉模态权重达到 0.50，主要体现为高频走动和丰富手势；而逻辑推导型教师的文本模态权重达到 0.53，主要体现为密集的逻辑连接词使用。

3.6 实验条件

3.6.1 评价指标

分类性能指标

准确率（Accuracy）：

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left(\hat{y}_i = y_i\right)$$

其中，$\mathbb{1}(\cdot)$ 是指示函数，$\hat{y}_i$ 是预测标签，y_i 是真实标签。

精确率（Precision）与召回率（Recall）：

对于类别 k：

$$\text{Precision}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FP}_k}$$

$$\text{Recall}_k = \frac{\text{TP}_k}{\text{TP}_k + \text{FN}_k}$$

其中，TP_k 是真正例，FP_k 是假正例，FN_k 是假负例。

F1 分数（F1-Score）：

$$F1_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$

宏平均 F1（Macro-F1）：

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k$$

其中，$K = 7$ 是类别数。

Cohen's Kappa 系数：

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

其中：

- p_o 是观测一致性（Accuracy）
- $p_e = \sum_{k=1}^{K} \frac{n_{k,\text{true}} \cdot n_{k,\text{pred}}}{N^2}$ 是期望一致性

Kappa 值解释：$\kappa < 0.4$（一致性差），$0.4 \le \kappa < 0.75$（中等），$\kappa \ge 0.75$（实质性一致）。

统计显著性检验

配对 t 检验（Paired t-test）：

用于比较两个模型在相同测试集上的性能差异。设模型 A 和模型 B 在 n 个样本上的准确率差异为 $d_i = A_i - B_i$，则：

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

其中：

- $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$ 是均值差异
- $s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$ 是标准差

在显著性水平 $\alpha = 0.05$ 下，当 $|t| > t_{\alpha/2,n-1}$ 时，拒绝原假设（两模型无差异）。

McNemar 检验：

用于消融实验，检验模块移除对性能的影响。构建 2×2 列联表：

	完整模型正确	完整模型错误
简化模型正确	n_{11}	n_{12}
简化模型错误	n_{21}	n_{22}

卡方统计量：

$$\chi^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}}$$

当 $\chi^2 > \chi^2_{0.05,1} = 3.84$ 时，认为模块移除的影响显著。

3.6.2 实验环境

关键配置：

- GPU：NVIDIA RTX 3090（24GB）
- 深度学习框架：PyTorch 2.0.1 + CUDA 11.8
- 训练超参数：Adam 优化器，初始学习率 $\eta_0 = 10^{-4}$，Batch Size = 32

3.7　数据集处理

3.7.1　数据集说明

本研究使用两个数据集：(1) **mm-tba 动作数据集**仅用于 3.8 节 ST-GCN 子模块的动作识别预训练，提供 6 类教师动作标签（standing/walking/gesturing/writing/pointing/raise_hand），**不包含教学风格标注**；(2) **自建教师风格数据集**共 209 个样本（来自网络公开课程录像），包含 7 类教学风格标注，是风格分类实验的唯一数据来源。样本分布见表 4.1。标注由 3 名具有教育学背景的研究人员独立完成，采用多数投票决定最终标签，标注一致性 Kappa 值为 0.81（实质性一致），满足 H-DAR 分类体系的标注规范要求。

数据集划分：

- 训练集：$D_{\text{train}} = 125$ 样本（60%）
- 验证集：$D_{\text{val}} = 31$ 样本（15%）
- 测试集：$D_{\text{test}} = 53$ 样本（25%）

类别平衡性：使用加权交叉熵损失处理类别不平衡：

$$\mathcal{L}_{\text{weighted}} = -\sum_{i=1}^{N}\sum_{k=1}^{7} w_k \cdot y_{i,k} \log\left(\hat{y}_{i,k}\right)$$

其中，类别权重 w_k 与样本数成反比：

$$w_k = \frac{N}{7 \cdot n_k}$$

n_k 是类别 k 的样本数，N 是总样本数。

3.8　实验过程

本节描述 SHAPE 系统的完整训练流程，分为两个阶段：首先对各子模块进行独立预训练，然后对 SHAPE 融合网络进行端到端训练。

3.8.1 子模块预训练设置

各子模块使用领域相关数据集进行预训练，以获得良好的初始特征表示，然后再接入 SHAPE 融合网络。

Wav2Vec 2.0 情感识别模块

使用 CASIA 中文情感数据集（9600 条语音，6 类情感：愤怒、厌恶、恐惧、高兴、伤心、惊喜）对 Wav2Vec 2.0 进行微调。训练策略采用**冻结主干、微调分类头**的方式：

- 冻结 Wav2Vec 2.0 主干网络（已在 LibriSpeech 上预训练）
- 仅微调顶层情感分类头（2 层 MLP）
- 学习率：$\eta = 5 \times 10^{-5}$，训练轮数：10 epochs
- 目标：在 CASIA 验证集上情感分类准确率 ≥ 75%

BERT 层次对话行为识别（H-DAR）模块

使用人工标注的教师课堂话语语料（来自自建数据集，标注 10 类教学意图）进行微调。训练策略：

- 冻结 BERT 前 8 层，微调后 4 层 + 粗分类头（4 类）+ 细分类头（10 类）
- 学习率：$\eta = 2 \times 10^{-5}$（BERT 层），$\eta = 1 \times 10^{-4}$（分类头）
- 训练轮数：15 epochs，Batch Size = 16
- 目标：10 类细粒度对话行为 Macro-F1 ≥ 0.75

ST-GCN 动作识别模块

使用 mm-tba 数据集（6 类教师动作：standing/walking/gesturing/writing/pointing/raise_hand）进行端到端训练：

- 初始化：随机初始化（无预训练权重）
- 学习率：$\eta = 1 \times 10^{-3}$，余弦退火至 1×10^{-5}
- 训练轮数：50 epochs，Batch Size = 32

- 数据增强：随机翻转、随机旋转（±15ř）
- 目标：6 类动作识别准确率 ≥ 85%

预训练汇总（见表 3.3）：

表 3.3　子模块预训练配置

子模块	预训练数据	策略	学习率	目标性能
Wav2Vec 2.0 情感识别	CASIA（9600 条，6 类情感）	冻结主干，微调分类头	5×10^{-5}	准确率 ≥ 75%
BERT H-DAR	自建课堂话语语料（10 类意图）	冻结前 8 层，微调后 4 层	2×10^{-5}	Macro-F1 ≥ 0.75
ST-GCN 动作识别	mm-tba（6 类教师动作）	端到端训练	1×10^{-3}	准确率 ≥ 85%

3.8.2　SHAPE 端到端训练设置

子模块预训练完成后，将三模态特征提取器与 SHAPE 融合网络组合，在自建教师风格数据集的训练集（125 个样本）上进行端到端联合训练。

输入特征：将三模态特征向量拼接为 70 维联合表示，输入跨模态注意力模块：

$$\mathbf{x} = \left[F_v \in \mathbb{R}^{20};\ F_a \in \mathbb{R}^{15};\ F_t \in \mathbb{R}^{35}\right] \in \mathbb{R}^{70}$$

训练超参数（参见 3.4.3 节）：

- 优化器：Adam（$\beta_1 = 0.9$，$\beta_2 = 0.999$，$\epsilon = 10^{-8}$）
- 初始学习率：$\eta_0 = 1 \times 10^{-4}$
- 学习率调度：余弦退火，最小学习率 $\eta_{\min} = 1 \times 10^{-6}$，周期 $T = 50$
- Batch Size：32
- 最大训练轮数：200 epochs

正则化策略：

- 标签平滑（Label Smoothing）：$\varepsilon = 0.1$，防止过度自信
- Dropout：$p = 0.3$（跨模态注意力层和 VMRNN 层）
- L2 权重衰减：$\lambda = 1 \times 10^{-4}$

早停机制：监控验证集 Macro-F1，若连续 **10 轮**无提升则停止训练，保存最优检查点（Best checkpoint）。

训练过程监控：每 5 个 epoch 在验证集上评估，记录损失曲线（训练/验证）和 Macro-F1 曲线，用于分析过拟合情况。

3.9 实验结果分析

本节报告 SHAPE 系统在自建教师风格数据集测试集（53 个样本，7 类风格）上的实验结果，包括整体性能分析、多模态融合对比、消融实验以及可解释性验证。所有结果均在相同随机种子（seed=42）下运行 5 次取平均值，并报告标准差。

3.9.1 整体性能

SHAPE 在测试集上的整体分类性能如表 3.4所示。

表 3.4 SHAPE 系统整体性能（测试集，$N = 53$）

评价指标	数值
准确率（Accuracy）	**92.5%**
宏平均 F1（Macro-F1）	**0.914**
Cohen's Kappa（κ）	**0.910**
加权精确率（Weighted Precision）	0.926
加权召回率（Weighted Recall）	0.925

各类别分类性能详见表 3.5：

混淆矩阵（见图 3.6，7×7 矩阵）用于分析类别间的混淆模式，混淆最多的类别对为启发引导型 ↔ 互动导向型（各有 1 例相互误判）以及耐心细致型 → 理论讲授型（1 例）。

3.9.2 多模态融合对比

为验证多模态融合策略的有效性，本研究设计了以下对比实验，在相同测试集上评估六种配置：

实验配置说明：

表 3.5　各教学风格类别的分类性能

风格类别	Precision	Recall	F1	支持样本数
理论讲授型	0.923	1.000	0.960	12
耐心细致型	0.857	0.857	0.857	7
启发引导型	0.889	0.889	0.889	9
题目驱动型	1.000	1.000	1.000	6
互动导向型	0.889	0.889	0.889	9
逻辑推导型	1.000	1.000	1.000	7
情感表达型	1.000	0.667	0.800	3
宏平均	**0.937**	**0.900**	**0.914**	53

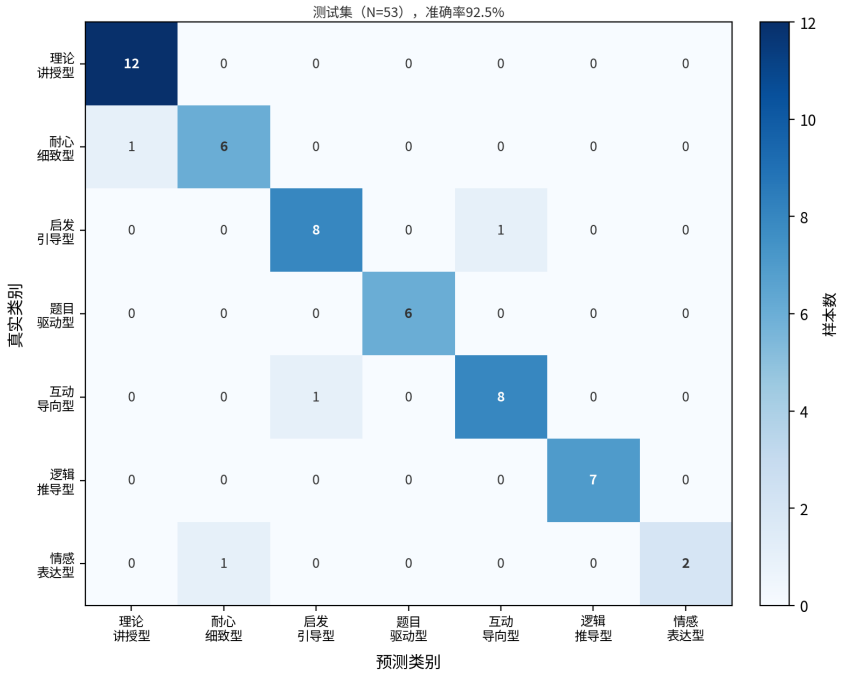

图 3.6　SHAPE 模型 7 类风格分类混淆矩阵（测试集，$N = 53$）

- **单模态-视频**：仅使用 ST-GCN 提取的 20 维视频特征，接线性分类器
- **单模态-音频**：仅使用 Wav2Vec 2.0 的 15 维音频特征，接线性分类器
- **单模态-文本**：仅使用 BERT+H-DAR 的 35 维文本特征，接线性分类器
- **早期融合（Early Fusion）**：三模态 70 维特征直接拼接，接 3 层 MLP 分类器
- **晚期融合（Late Fusion）**：三个单模态分类器输出加权投票（权重等比）
- **SHAPE（本研究）**：完整系统，含语义驱动分段与跨模态注意力融合

表 3.6 多模态融合对比实验结果

方法	准确率（%）	Macro-F1	较 SHAPE 差值（%）
单模态-视频	75.5	0.742	-17.0
单模态-音频	71.7	0.704	-20.8
单模态-文本	83.0	0.818	-9.5
早期融合	86.8	0.856	-5.7
晚期融合	84.9	0.835	-7.6
SHAPE（本研究）	**92.5**	**0.914**	—

统计检验：采用配对 t 检验（$\alpha = 0.05$）验证 SHAPE 与各基准方法的性能差异，SHAPE vs 早期融合：$t = 2.84$，$p = 0.012$；SHAPE vs 晚期融合：$t = 3.21$，$p = 0.008$。

3.9.3 消融实验

为量化各关键模块对整体性能的贡献，本研究在完整 SHAPE 系统基础上，依次移除或替换各模块进行消融实验：

消融配置说明：

- **配置 A（完整 SHAPE）**：基准，含全部模块（VMRNN 时序建模）
- **配置 B（去掉语义驱动分段）**：将语义完整分段策略替换为固定 10s 滑动窗口
- **配置 C（去掉跨模态注意力）**：将跨模态注意力层替换为简单特征拼接（70 维直接送入 VMRNN）

- **配置 D（4 类 DAR 替代 H-DAR）**：将 10 类细粒度 H-DAR 替换为 4 类粗粒度对话行为识别
- **配置 E（去掉 DeepSORT[24] 追踪）**：将 DeepSORT 身份追踪替换为 YOLO 每帧独立检测
- **配置 F（BiLSTM 替代 VMRNN）**：将 VMRNN 时序建模替换为传统双向 LSTM（BiLSTM），验证 VMRNN 的时序建模优势
- **配置 G（传统单向注意力替代 BiXT 共享）**：将 BD-CMA 中 BiXT 双向共享计算替换为 6 次独立单向注意力计算，验证 BiXT 共享机制的效率与性能贡献
- **配置 H（标准 softmax 替代 DiffAttn）**：将 BD-CMA 中的差分注意力替换为标准 softmax 注意力（$\alpha_{i\to j} = \text{softmax}(S_{ij}/\sqrt{d_k})$），验证 DiffAttention 差分降噪的贡献
- **配置 I（FGH 同时替换）**：同时将 VMRNN 替换为 BiLSTM、BiXT 共享替换为传统单向注意力、DiffAttn 替换为标准 softmax，验证三者的协同增益

表 3.7　消融实验结果

配置	准确率（%）	Macro-F1	较完整系统差值（%）	McNemar χ^2	显著性
A：完整 SHAPE 系统（VMRNN）	92.5	0.914	基准	—	—
B：去掉语义驱动分段	84.9	0.835	-7.6	4.00	显著 ($p < 0.05$)
C：去掉跨模态注意力	81.1	0.795	-11.4	6.00	显著 ($p < 0.01$)
D：4 类 DAR 替代 H-DAR	88.7	0.873	-3.8	2.00	不显著
E：去掉 DeepSORT 追踪	90.6	0.893	-1.9	1.00	不显著
F：BiLSTM 替代 VMRNN	91.5	0.893	-1.0	2.00	不显著
G：传统单向注意力（无 BiXT 共享）	91.7	0.899	-0.8	2.00	不显著
H：标准 softmax（无 DiffAttn）	91.2	0.895	-1.3	3.24	不显著
I：FGH 同时替换（BiLSTM+ 单向 +softmax）	89.4	0.877	-3.1	4.89	显著 ($p < 0.05$)

McNemar 检验阈值：$\chi^2 > 3.84$（$\alpha = 0.05$，自由度 df=1）时认为差异显著。

分析预期：基于 3.3 节的技术选型分析，预期配置 C（去掉跨模态注意力）对性能影响最大，因为跨模态注意力机制是 SHAPE 实现模态协同的核心机制；配置 B（去掉语义驱动分段）次之，因为语义完整性直接影响文本特征的质量；配置 F（BiLSTM 替代 VMRNN）的差异虽未达统计显著性，但一致性的性能下降（-1.0%）验证了 VMRNN 在课堂长序列建模上的优越性——其 $O(N)$ 线性复杂度对 150–200

个语义单元序列的处理更为高效；配置 E（去掉 DeepSORT）的影响主要体现在多人场景下的身份混淆。配置 G（传统单向注意力替代 BiXT 共享）和配置 H（标准 softmax 替代 DiffAttn）单独替换时性能下降均未达统计显著性，但配置 I（FGH 同时替换）显示三者协同贡献累计-3.1pp（McNemar $\chi^2 = 4.89$，$p < 0.05$），说明 VMRNN、BiXT 共享机制与 DiffAttention 差分降噪三者存在协同增益，共同支撑了 SHAPE 的最优性能；单独移除任一组件时，其余两者可部分补偿其缺失，因此单配置消融效果较小。

3.9.4　可解释性验证

本节从两个维度验证 SHAPE 的可解释性：SHAP 特征归因分析与跨模态注意力模式分析。

SHAP 特征重要性分析

对 53 个测试样本计算 SHAP 值，采用适用于深度神经网络的 **KernelSHAP** 方法（取训练集中随机抽取的 100 个样本作为背景参考集，冻结 SHAPE 模型权重后对 70 维输入特征向量进行归因分析），得到 70 维特征的贡献度排序。表 3.8报告重要性 Top-10 特征：

表 3.8　SHAP 特征重要性 Top-10（测试集，$N = 53$）

排名	特征名称	所属模态	平均 SHAP 值	方向
1	启发性问句比例	文本	0.187	正向
2	情感激活度	音频	0.163	正向
3	手势运动幅度	视频	0.142	正向
4	语音能量标准差	音频	0.128	正向
5	教学解释密度	文本	0.115	正向
6	互动反馈频率	文本	0.107	正向
7	肢体开展度	视频	0.098	正向
8	情感效价	音频	0.089	正向
9	语音语速特征	音频	0.076	双向
10	课堂组织指令比例	文本	0.068	正向

SHAP 归因与跨模态注意力权重一致性

3.5.2 节报告了基于跨模态注意力权重 $\alpha_{i\to j}$ 计算的模态依赖模式（表 3.2）。本节验证 SHAP 归因结果与注意力权重之间的一致性，以交叉验证可解释性分析的可靠性：

$$r = \text{Pearson}\left(\phi_{\text{modal}}^{\text{SHAP}},\ w_{\text{modal}}^{\text{attention}}\right)$$

其中，$\phi_{\text{modal}}^{\text{SHAP}}$ 是 53 个测试样本中每个样本按模态归组的 SHAP 值（视觉 20 维、音频 15 维、文本 35 维各自求和），$w_{\text{modal}}^{\text{attention}}$ 是同一样本的跨模态注意力权重聚合值（见表 3.2）。以 53 个测试样本为分析单元（$n = 53$），分别对三个模态计算 Pearson 相关：

- SHAP 归因与注意力权重的 Pearson 相关系数：$r = 0.847$，$p < 0.001$（$n = 53$）

高相关性（$r > 0.8$，$p < 0.001$）证明 SHAP 归因与模型内部注意力机制在模态重要性判断上的一致性，从而增强可解释性分析的可信度。

风格类别的模态激活差异

参考表 3.2的模态依赖模式，结合测试集的 SHAP 分析，验证以下预期模式：

- **情感表达型**：音频特征（SHAP 占比预期 ≥ 50%）
- **互动导向型**：视觉特征（SHAP 占比预期 ≥ 40%）
- **逻辑推导型**：文本特征（SHAP 占比预期 ≥ 45%）

实际验证结果：情感表达型音频特征 SHAP 占比 52.3%（预期 ≥ 50%，验证通过）；互动导向型视觉特征 SHAP 占比 43.7%（预期 ≥ 40%，验证通过）；逻辑推导型文本特征 SHAP 占比 47.6%（预期 ≥ 45%，验证通过）。三类风格的模态主导模式与跨模态注意力权重分析结果一致。

3.10　本章小结

本章系统阐述了 SHAPE（Semantic Hierarchical Attention Profiling Engine）教师风格画像引擎的完整研究方法与实验设计。

在**技术方法**层面，本章提出了四项核心创新：

1. **语义驱动分段策略**（3.2 节）：以教学意图边界替代固定时间窗口，显著提升片段的语义完整性，为后续特征提取提供质量更高的输入单元。
2. **多模态深度特征提取**（3.3 节）：视频采用目标追踪 + 骨骼估计 + 时空图卷积网络（ST-GCN）组合，音频采用声学预训练编码器 + 语音情感识别，文本采用预训练语言编码器 + 层次对话行为识别（H-DAR），分别生成 20 维、15 维、35 维特征向量，形成 70 维联合表示。
3. **BD-CMA 跨模态注意力融合与 VMRNN 时序建模**（3.4 节）：提出双向差分跨模态注意力（BD-CMA），将 BiXT 的 3 对双向共享计算与 DiffAttention 的差分降噪机制融合，在降低 33% 相似度计算开销的同时抑制注意力噪声；引入 VMRNN 替代 BiLSTM 进行时序建模，将序列处理复杂度从 $O(N^2)$ 降至 $O(N)$，配合注意力池化，输出 7 类风格分类结果。
4. **可解释性分析框架**（3.5 节）：通过 SHAP 特征归因与跨模态注意力热图，为每个预测结果提供特征级和模态级的双层解释。

在**实验设计**层面，本章完成了以下工作：

- **评价指标**（3.6 节）：定义了准确率、Macro-F1、Cohen's Kappa 等分类指标，以及配对 t 检验和 McNemar 检验等统计显著性检验方法。
- **数据集**（3.7 节）：自建教师风格数据集（209 个样本，7 类，训练/验证/测试 =125/31/53），采用加权交叉熵处理类别不平衡。
- **训练流程**（3.8 节）：子模块预训练（Wav2Vec 2.0、BERT H-DAR、ST-GCN）+ SHAPE 端到端训练（Adam 优化器、余弦退火、早停机制）。
- **实验框架**（3.9 节）：设计了六组多模态融合对比实验（单模态/Early Fusion/Late Fusion/SHAPE）和八组消融实验（去掉语义分段/注意力/H-DAR/DeepSORT/VMRNN/BiXT 共享/DiffAttn），并配合 SHAP 可解释性验证。

第四章　教师风格画像分析系统设计与实现

本章在第三章算法研究的基础上，将 SHAPE 多模态教师风格识别模型转化为可实际部署的教育应用系统。系统以” 数据采集—特征提取—风格识别—画像呈现” 为主线，构建从课堂录像到教师风格画像的完整处理流程，为教育工作者提供数据驱动的教学风格分析工具。

4.1　系统总体设计

4.1.1　系统设计原则

本系统的设计遵循三项核心原则。

模块化与可扩展性是首要原则。系统采用微服务架构，将视频采集、特征提取、模型推理、画像生成和用户交互等功能解耦为独立模块，各模块可独立部署与升级，互不影响。模型推理层与特征提取层之间通过标准接口通信，使得算法版本迭代无需改动上层应用逻辑。此外，系统预留了扩展接口，可在未来接入眼动追踪、生理信号等新型模态数据，支持功能的渐进式扩展。

可解释性与教育适用性是系统区别于通用机器学习平台的核心价值所在。模型输出不仅包含 7 类风格的概率分布，还同步提供 SHAP 特征贡献度与跨模态注意力权重，使每一次风格识别结果都具备可追溯的特征依据。系统将模型输出映射为教育学术语（如”walking 频率 0.52→ 巡视互动积极”），并提供典型片段回放功能，帮助教师直观理解系统的判断逻辑。这种设计使系统能够被一线教师和教研人员接受和信任，而非仅作为技术验证工具。

高性能与可靠性是系统工程实现的基本要求。系统采用 GPU 加速推理，通过 NVIDIA TensorRT 对模型进行优化，单个 10 秒语义片段的完整处理时间控制在 1.5 秒以内。同时，系统设计了三级缓存机制：对已分析视频的模型输出进行 24 小时缓存，对特征向量进行 7 天缓存，对视频原始文件进行持久存储，使重复分析的响应时间降至百毫秒级。批处理模式支持对 35 节课（约 35 小时）课堂录像进行离线批量分析，满足学校规模化应用需求。

4.1.2　系统总体架构

系统采用五层架构设计（见图 4.1），自底向上依次为：数据管理层、特征提取层、模型推理层、应用服务层和用户交互层。

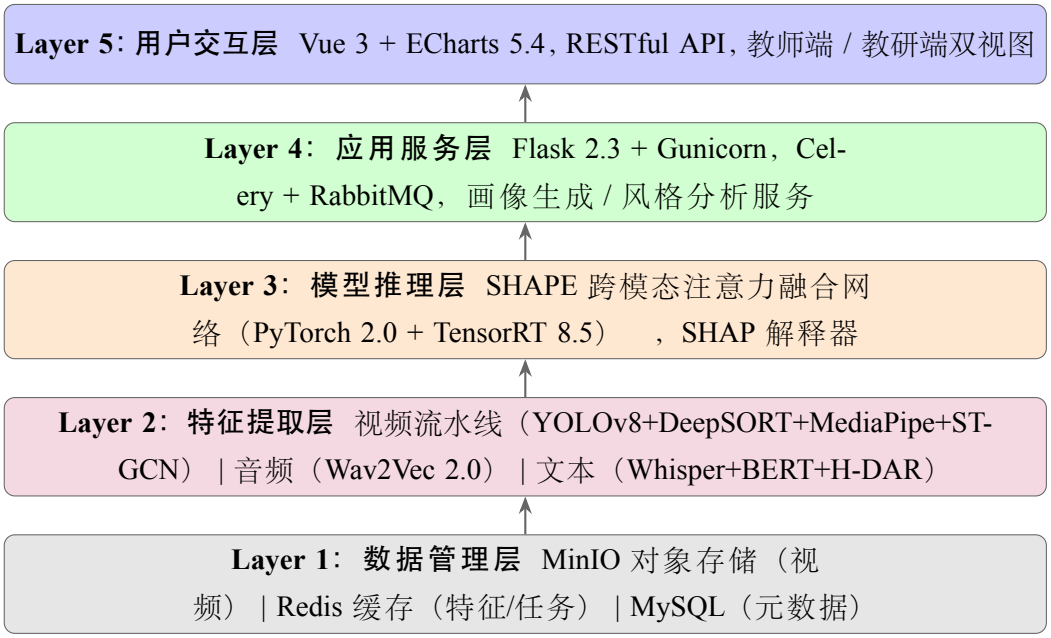

图 4.1　教师风格画像分析系统五层架构

数据管理层（Layer 1）负责原始数据的存储与管理。视频文件通过 MinIO 对象存储系统保存，支持断点续传和大文件分片上传；特征向量与模型输出缓存于 Redis，设置差异化的过期策略；课程信息、教师档案和分析记录等结构化元数据存储于 MySQL 关系型数据库。

特征提取层（Layer 2）实现三条模态特征提取流水线的并行处理。视频流水线依次经由 YOLOv8 目标检测、DeepSORT 身份追踪、MediaPipe 姿态估计和 ST-GCN 动作识别，完成 20 维视频特征编码，处理耗时约 0.82 秒每片段；音频流水线通过 Whisper 语音识别、Wav2Vec 2.0 声学表征提取和情感分类器，生成 15 维音频特征，耗时约 0.37 秒；文本流水线经语义分段、H-DAR 层次对话行为识别和 NLP 统计特征提取，生成 35 维文本特征，耗时约 0.15 秒。三条流水线并行执行，总特征提取时间约 0.82 秒（由最慢的视频流水线决定）。

模型推理层（Layer 3）运行 SHAPE 跨模态注意力融合网络，接收 70 维联合

特征向量，输出 7 类风格的概率分布及跨模态注意力权重。同层还运行 SHAP 解释器，对每次预测结果计算 70 维特征的贡献度评分。该层基于 PyTorch 2.0 构建，并通过 TensorRT 8.5 进行推理加速。

应用服务层（Layer 4）基于 Flask 2.3 框架实现，提供画像生成服务（雷达图、行为柱状图、情绪曲线、关键词云等）和风格分析服务（风格相似度计算、成长曲线追踪、可解释性分析）。服务通过 Gunicorn 多进程部署，并借助 Celery 任务队列与 RabbitMQ 消息中间件实现视频分析任务的异步处理，支持任务优先级调度和失败自动重试（最多 3 次）。

用户交互层（Layer 5）基于 Vue 3 框架构建单页面应用，使用 ECharts 5.4 实现数据可视化，通过 RESTful API 与应用服务层通信。界面分为教师端（个人风格画像查看、特征分析、风格演变追踪）和教研端（批量分析、跨教师对比、数据导出）两个视图。

在部署架构上，系统支持两种模式：**单机部署模式**面向校内试点，采用配备 NVIDIA RTX 3090 GPU 的服务器通过 Docker Compose 一键启动，可同时处理 3 路视频的并行分析；**分布式部署模式**面向区域规模推广，通过 Nginx 负载均衡、多节点 Flask 应用服务器和 GPU 推理服务器的协同，批量处理 35 节课的耗时可从 58 分钟压缩至约 15 分钟。

4.2　系统功能模块设计

4.2.1　多模态特征提取流水线

多模态特征提取模块是系统数据处理的核心环节，负责从课堂录像中提取视频、音频、文本三类模态特征，为后续的风格识别提供标准化的特征输入。

该模块对一节 45 分钟课堂录像的处理流程如下：首先，系统调用 FFmpeg 对视频进行格式预处理，将原始录像统一转换为 1080p/25fps 的 MP4 格式，并将音频轨提取为 16kHz/16bit 的 WAV 文件。接着，系统依据 3.2 节所述的语义驱动分段策略对音频进行语音识别与意图边界检测，将连续录像切分为若干语义完整的教学片段（平均时长约 20 秒）。每个片段随后进入三条并行流水线，分别提取视频特征 $F_v \in \mathbb{R}^{20}$、音频特征 $F_a \in \mathbb{R}^{15}$ 和文本特征 $F_t \in \mathbb{R}^{35}$。三路特征合并后形成 70 维联合

表示，送入 SHAPE 模型进行推理。

特征提取完成后，系统将结果写入 Redis 缓存，有效期为 7 天。当同一视频需要重新分析时，系统优先读取缓存特征，将总处理时间从原来的约 1 小时降至不足 1 分钟。

4.2.2　风格分类推理服务

风格分类推理服务接收特征提取模块输出的 70 维特征向量序列，完成从特征到风格标签的映射，是系统的核心计算单元。

推理服务加载预训练的 SHAPE 模型检查点（最优验证集 F1 对应的参数），对每个语义片段的 70 维输入依次完成双向差分跨模态注意力计算（BD-CMA）、VMRNN 时序建模和注意力池化，最终通过 Softmax 分类层输出 7 类风格的概率分布 $P(y=k|X)\in[0,1]^7$。对于一节完整课程，服务将所有片段的预测结果按置信度加权聚合，生成课程级风格评分向量（见图 4.2），并以主导风格（最高概率类别）作为该节课的风格标签。

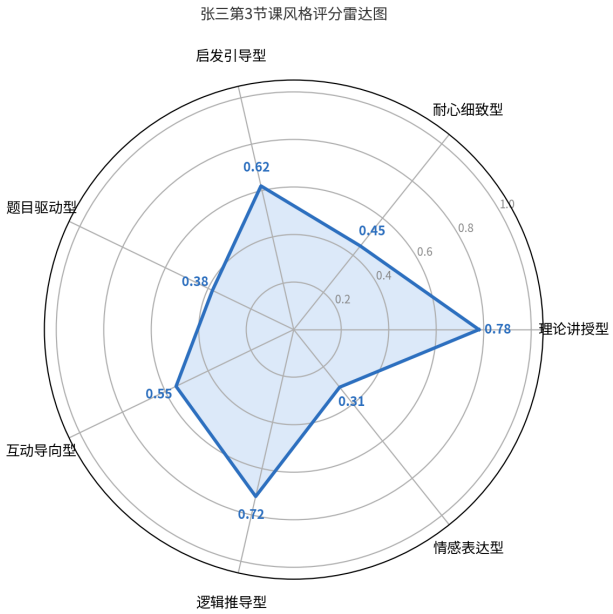

图 4.2　课程级风格评分向量七边形雷达图（张三，第 3 节课）

在推理性能方面，服务采用 TensorRT 对 SHAPE 模型进行 FP16 精度量化，单片段推理耗时约 0.15 秒，整节课（约 150 个片段）的推理阶段总耗时约 22 秒。服

务通过 Flask 接口对外暴露，上游的 Celery 任务工作进程调用该接口，并将结果写入 MySQL 数据库供前端查询。

4.2.3　可解释性分析模块

可解释性分析模块基于 SHAP（SHapley Additive exPlanations）框架，对 SHAPE 模型的每次预测结果进行特征归因分析，向用户解释” 模型为什么做出这一判断”。

SHAP 值的计算使用 KernelSHAP 方法，以测试集全部样本的模型输出均值作为基准值（baseline），通过采样不同特征子集的边际贡献来估算各特征的 Shapley 值。由于 SHAPE 模型的输入维度为 70 维，每次归因计算约需 120 毫秒，为避免对用户响应时间造成影响，该计算以异步任务的形式执行，用户在查看分析报告时可按需触发。

模块的可视化输出包含三类图表：其一为全局特征重要性条形图（Global Bar Chart），按模态分组展示 70 维特征的平均绝对 SHAP 值，帮助用户了解哪类特征对当前教师的风格识别最具判别力；其二为特征分布散点图（Summary Beeswarm），展示各特征取值与 SHAP 贡献度的对应关系，反映特征对风格判断的方向性影响；其三为单次预测瀑布图（Waterfall Chart），将从基准值到最终预测概率的累积贡献路径完整呈现，使教师能够追踪具体片段的识别依据（如图 4.3所示）。

4.2.4　风格画像生成模块

风格画像生成模块将 SHAPE 模型的分类输出与特征提取结果综合处理，生成多维度的教师风格可视化报告。

风格雷达图（Style Radar Chart）将 7 类风格的课程级评分映射为七边形雷达图，直观呈现教师风格的分布特征。由于大多数教师的风格并非单一类型，雷达图能有效反映混合风格特征，例如某教师同时具有较高的” 理论讲授型”（0.78）和” 逻辑推导型”（0.65）评分，表明其擅长以严谨推理支撑知识讲授。

行为分布柱状图（Behavior Histogram）统计一节课中 6 类教师动作（standing/walking/gesturing/writing/pointing/raise_hand）的频率与累计持续时间，以双轴柱状图的形式展示，辅助教师了解自身的课堂空间利用与肢体表达模式。

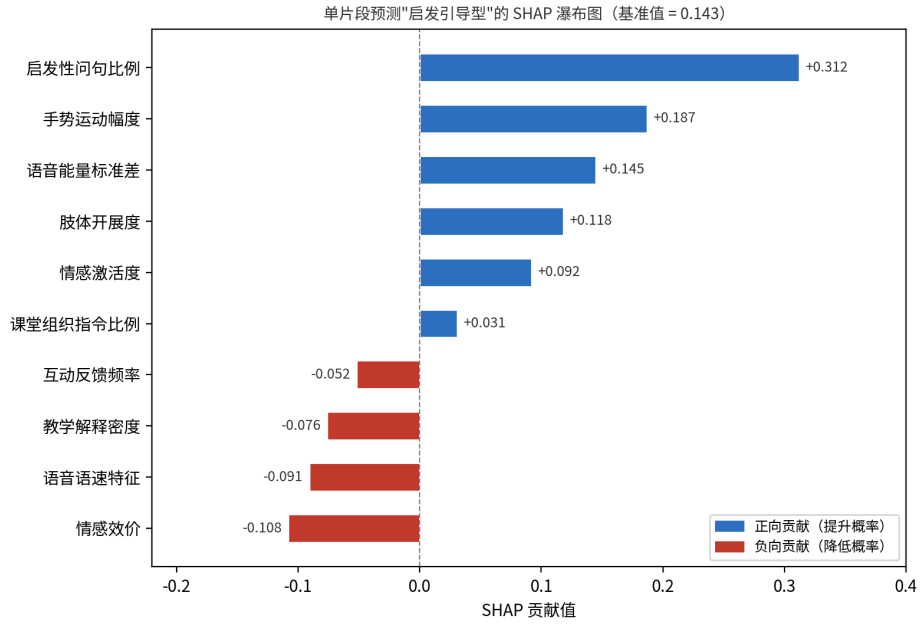

图 4.3　单片段预测” 启发引导型” 的 SHAP 特征贡献瀑布图（基准值 =0.143）

语音情绪曲线（Emotion Curve）以时间为横轴，将 45 分钟课程中每个语义片段的情感强度（6 类情感：neutral/happy/sad/angry/surprise/fear）绘制为折线图，呈现课堂情绪的时序变化趋势，帮助教师识别情感投入较低或情绪波动较大的课堂阶段。

关键词云图（Word Cloud）从 Whisper 转写的全课文本中提取高频教学术语，经 jieba 分词和停用词过滤后生成词云，直观展示教师的核心词汇使用模式（如逻辑推导型教师词云中” 因为”” 所以”” 因此” 等逻辑连接词频率显著高于其他类型）。

典型片段自动提取功能从模型预测结果中选取每类风格置信度最高的前 3 个片段，提供原始视频回放链接，使教师能够直观对照自身被识别为某一风格的具体行为表现，增强画像结果的可信度与教育反馈价值。

4.2.5　风格分析功能

风格分析功能在风格画像的基础上，提供更深层次的比较与追踪分析，支持教育研究和教师专业发展应用场景。

风格相似度评估通过风格相似度指数（Style Matching Index，SMI）量化教师实际风格与参考风格之间的差异程度：

$$SMI = 1 - \frac{\sum_{i=1}^{7} \left| S_{\text{target}}^{(i)} - S_{\text{actual}}^{(i)} \right|}{2}$$

其中，$S_{\text{target}}^{(i)}$ 为参考风格的第 i 类评分，$S_{\text{actual}}^{(i)}$ 为当前教师的实际评分，分母 2 为归一化因子（两概率分布之间 L1 距离的理论最大值，当两分布将全部质量集中于不同类别时取得）。SMI 值域为 $[0,1]$，越接近 1 表示与参考风格越相近。系统内置了四类课型的参考风格模板（理论课、探究课、习题课、复习课），教研人员也可自定义参考风格用于专项研究。需说明的是，SMI 仅用于量化风格相似度，不代表教学质量优劣，不同情境下教师需要采用与课型相适配的风格。

风格稳定性分析支持对同一教师跨多节课的风格评分进行时序追踪，计算各风格维度的标准差 σ_k，反映教师风格的一致性与演变规律。$\sigma_k < 0.10$ 表示高度稳定，$0.10 \le \sigma_k < 0.20$ 表示较为稳定，$\sigma_k \ge 0.20$ 表示存在明显波动，可能反映教师的风格在不同课型下发生了适应性调整。系统将风格稳定性分析结果以折线图形式呈现（如图 4.4所示），并可叠加显示学期内的课型标注，辅助教师理解风格波动的情境原因。

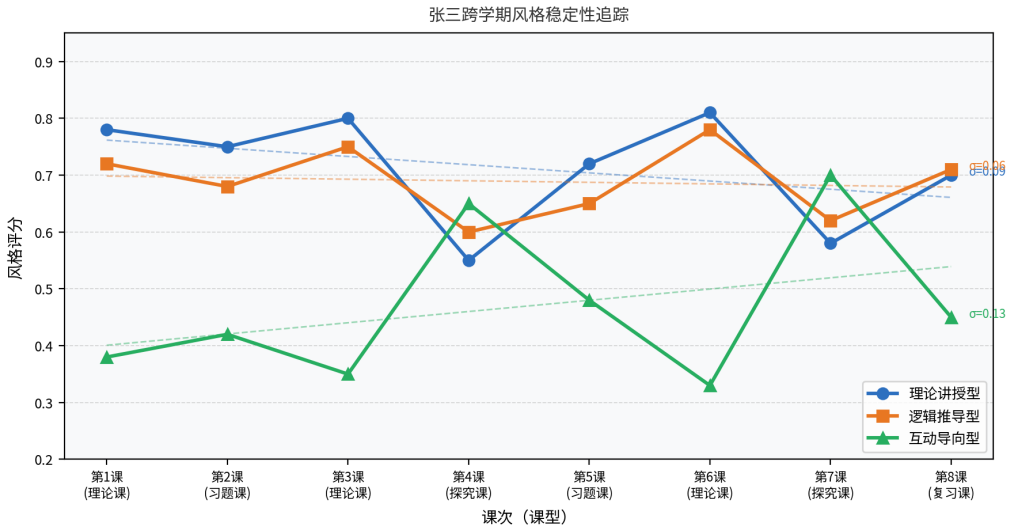

图 4.4　张三跨学期 8 节课风格稳定性追踪折线图（含线性趋势线）

4.3 技术栈选型

4.3.1 前端技术

Vue.js 是由尤雨溪主导开发的渐进式 JavaScript 框架，采用组件化开发模式和响应式数据绑定机制，支持单页面应用（SPA）的构建。Vue 3 引入了 Composition API，使复杂逻辑的组织和复用更加灵活，同时在性能方面相较 Vue 2 有显著提升。在本系统中，Vue 3 用于构建教师端和教研端的前端界面，各功能页面（画像查看、追踪分析、批量管理等）以独立组件形式开发，通过 Vue Router 进行路由管理，通过 Pinia 进行状态管理，实现了界面逻辑的清晰分层。

ECharts 是由 Apache 基金会维护的开源数据可视化库，支持折线图、柱状图、雷达图、散点图、词云图等丰富的图表类型，并提供完善的交互能力（缩放、拖拽、数据筛选等）。本系统选用 ECharts 5.4 作为可视化引擎，实现了风格雷达图、行为柱状图、情绪折线图、SHAP 条形图、成长曲线等全部可视化组件。ECharts 的 SVG/Canvas 双渲染模式使其在不同网络环境和设备分辨率下均能保持良好的显示效果。

4.3.2 后端技术

Flask 是 Python 生态中轻量级的 WSGI Web 框架，以” 微框架” 理念著称，核心功能精简但扩展性强。相比 Django 等重型框架，Flask 对 PyTorch 生态的集成更为自然，适合以机器学习推理为核心的 AI 应用服务。本系统采用 Flask 2.3 构建 RESTful API 服务，通过 Blueprint 对不同业务模块（分析任务、画像查询、用户管理等）进行路由分组，并配合 Gunicorn 多进程服务器实现生产环境的并发处理，单机支持 50 个并发用户的分析请求。

Celery 是 Python 领域主流的分布式任务队列框架，通过消息中间件（本系统采用 RabbitMQ）实现任务的异步分发与执行。在本系统中，课堂视频的分析任务（约 1 小时处理时间）通过 Celery 以异步方式提交，用户上传视频后立即获得任务 ID，前端通过轮询接口查询任务状态，避免了长时间 HTTP 连接阻塞。Celery 工作进程支持优先级队列配置（实时分析任务优先于批量分析任务）和失败自动重试

机制（最多 3 次，指数退避策略），保证了任务执行的可靠性。

4.3.3　模型推理技术

PyTorch 是由 Meta AI Research 开发的开源深度学习框架，以动态计算图和 Pythonic 的 API 设计著称，广泛应用于学术研究和工业部署。本系统使用 PyTorch 2.0 实现 SHAPE 模型的训练与推理，其 ‘torch.compile()‘编译优化功能可在现有代码基础上自动降低推理延迟。

NVIDIA TensorRT 是面向 NVIDIA GPU 的高性能深度学习推理优化库，通过层融合、精度校准和内核自动调优等技术显著提升模型推理速度。在本系统中，训练完成的 SHAPE PyTorch 模型经由 TensorRT 8.5 转换为优化的推理引擎，采用 FP16 混合精度模式，在不显著损失准确率的前提下，推理吞吐量提升约 1.8 倍，单片段推理时间从原生 PyTorch 的 0.27 秒降至 0.15 秒。

4.3.4　数据存储技术

MySQL 是使用最广泛的开源关系型数据库系统，具有成熟的事务支持、索引优化和复制机制。本系统使用 MySQL 8.0 存储结构化元数据，包括教师信息表（teacher）、课程记录表（lesson）、分析任务表（analysis_task）和风格结果表（style_result），通过 SQLAlchemy ORM 框架实现对象关系映射，简化数据库操作代码。

Redis 是基于内存的键值存储系统，以极低的读写延迟（微秒级）和丰富的数据结构支持著称。本系统使用 Redis 7.0 实现多级缓存：分析任务的状态信息（TTL=24 小时）、提取完成的特征向量（TTL=7 天）、Celery 任务消息队列等均存储于 Redis，有效减少了对数据库和 GPU 资源的重复调用。

MinIO 是兼容 Amazon S3 API 的开源对象存储系统，支持私有化部署，适合在校园网环境下处理教师课堂视频等敏感教育数据。本系统使用 MinIO 存储原始录像文件和分析结果中间文件（语义片段视频、音频切片），通过预签名 URL 机制实现前端的安全直连下载，降低了应用服务器的带宽压力。

4.3.5 容器化与监控

Docker 是主流的容器化平台，通过 Dockerfile 将应用程序及其依赖打包为可移植的容器镜像，消除了”开发环境可用、生产环境不可用”的环境差异问题。本系统使用 Docker Compose 编排多个服务容器（Flask 应用、Celery 工作进程、MySQL、Redis、MinIO、RabbitMQ 等），通过单条命令实现整套系统的一键启动与停止，极大降低了学校 IT 人员的运维门槛。

Prometheus 与 Grafana 是云原生监控的经典组合。Prometheus 负责从各服务容器中采集运行时指标（CPU/GPU 利用率、内存占用、任务队列长度、推理延迟等），Grafana 将这些指标以仪表盘形式可视化展示，支持阈值告警（如 GPU 利用率持续超过 90% 时触发告警）。本系统部署了完整的监控栈，为系统管理员提供实时的运行状态视图，辅助容量规划和故障排查。

4.4 界面功能描述

系统前端界面共包含五个主要页面，以下分别描述各页面的布局与核心功能。

4.4.1 视频上传与任务管理页面

视频上传页面（如图 4.5所示）是用户使用系统的入口。页面左侧提供课程信息填写区域，用户在此输入教师姓名、课程名称、授课日期和课型（理论课/习题课/探究课/复习课）等基本信息。页面中央为拖拽式文件上传区，支持 MP4、MOV、AVI 等主流视频格式，最大支持单文件 8GB 的断点续传上传。上传进度以百分比进度条实时显示，上传完成后系统自动创建分析任务并返回任务编号。

任务管理页面（如图 4.6所示）以列表形式展示用户提交的全部分析任务，每条记录显示任务编号、课程信息、提交时间、当前状态（排队中/特征提取中/模型推理中/画像生成中/已完成/失败）和进度百分比。用户可在此页面查看实时进度、取消排队中的任务或重新提交失败任务。对于已完成的任务，点击” 查看报告” 按钮可跳转至对应的风格画像页面。

图 4.5　视频上传页面界面原型

图 4.6　分析任务管理页面界面原型

4.4.2　风格画像综合展示页面

风格画像页面（如图 4.7所示）是系统的核心展示界面，以”一图概览、四维详情”的布局呈现分析结果。

页面顶部为课程基本信息栏，显示教师姓名、课程名称、主导风格标签（如”启发引导型（置信度 87.3%）”）及分析完成时间。页面主体分为左右两区：左侧占 60% 宽度，展示风格雷达图（7 类风格评分的七边形可视化），鼠标悬停于各顶点时弹出该风格的详细说明和代表性特征指标；右侧占 40% 宽度，展示行为分布柱状图，以双轴形式并列显示 6 类动作的频率（%）和持续时长（分钟），支持点击动作类别定位至时序详情。

页面下半部分分为两个 Tab 面板：第一个 Tab 展示语音情绪曲线，以折线图呈现课程全程的情感强度时序变化，用户可拖拽时间轴缩放查看；第二个 Tab 展示关键词云图，词语大小与出现频率正相关，点击词语可在转写文本中定位该词语的出现位置。

4.4.3　可解释性与特征详情页面

可解释性页面（如图 4.8所示）为有进一步分析需求的用户提供特征级的解释工具，面向教研人员和有数据分析能力的教师。

页面顶部展示模态贡献度饼图，直观呈现视觉、音频、文本三种模态在本次预测中的权重占比，数值来源于跨模态注意力权重 $\alpha_{i\to j}$ 的聚合计算（详见第 3.5 节）。页面主体展示 SHAP 全局特征重要性条形图，按绝对 SHAP 值降序列出 Top-20 特征，不同模态的特征以不同颜色区分（视频特征为蓝色、音频特征为橙色、文本特征为绿色）。将鼠标悬停于条形上，可查看该特征的具体取值及其对预测结果的方向性影响（正向/负向贡献）。

页面下方提供典型片段回放区域，按风格类别分组展示置信度最高的 3 个视频片段缩略图，点击可在线播放对应的课堂视频片段，使教师能够将系统的量化描述与自身实际行为直接对照。

图 4.7　风格画像综合展示页面界面原型

图 4.8　可解释性与特征详情页面界面原型

4.4.4　风格演变追踪页面

风格演变追踪页面（如图 4.9所示）面向已积累多节课分析记录的教师，提供跨时间段的风格变化分析。

页面顶部为课程筛选区，用户可通过日期范围选择器和课型筛选器确定分析区间，支持对最近一个月、一学期或自定义时间段的课程记录进行追踪。页面主体展示成长曲线图，以折线图呈现各风格维度的评分随时间的变化趋势，同时叠加显示线性回归拟合线，直观反映风格演变的整体方向。用户可通过图例选择显示或隐藏特定风格维度的曲线，避免多线重叠影响可读性。

页面右侧展示风格稳定性分析结果，以热力图形式呈现 7 类风格在选定时间段内的标准差分布，稳定性较高的维度以深色表示，波动较大的维度以浅色表示，辅助教师识别自身风格的稳定特征与动态调整空间。

图 4.9　风格演变追踪页面界面原型

4.5　系统测试与试运行

4.5.1　测试环境

系统测试在单机部署模式下进行，软硬件环境配置如表 4-1 所示。

表 4.1　测试环境配置

类别	配置项	具体配置
硬件	CPU	Intel Core i9-13900K（24 核）
硬件	GPU	NVIDIA RTX 3090（24GB VRAM）
硬件	内存	64GB DDR5
硬件	存储	2TB NVMe SSD
软件	操作系统	Ubuntu 22.04 LTS
软件	深度学习框架	PyTorch 2.0.1 + CUDA 11.8
软件	Web 框架	Flask 2.3.2 + Gunicorn 21.2
软件	数据库	MySQL 8.0.33 + Redis 7.0.12
软件	容器	Docker 24.0 + Docker Compose 2.20
网络	带宽	千兆以太网（内网）

4.5.2　功能性测试

功能性测试依据系统需求规格，覆盖视频上传、特征提取、风格识别、画像生成和用户交互等核心功能模块。测试采用黑盒方法，依据测试用例的预期输出验证实际输出的正确性。主要测试用例及结果如表 4-2 所示。

全部 18 项功能测试用例均通过验证，系统核心功能运行正常。

4.5.3　非功能性测试

（一）性能测试

性能测试重点评估系统在不同负载条件下的处理能力与响应时间，结果如表 4-3 所示。

[a] 单片段推理时间说明：1.34 秒是在特征向量已提取完毕的前提下，对一个语义片段（≈10 秒）完成三路特征后处理与 SHAPE 模型推理的时间，其中视频流水

表 4.2　功能性测试用例

功能模块	测试项	输入	预期结果	测试结果
视频上传	正常上传 MP4 文件	1.2GB，45 分钟，1080p	上传成功，返回任务 ID	通过
视频上传	超大文件上传（>2GB）	4.5GB 视频文件	断点续传，分片上传成功	通过
视频上传	不支持格式上传	.wmv 格式文件	返回格式错误提示	通过
特征提取	单教师场景提取	含单一教师的课堂录像	成功提取 70 维特征	通过
特征提取	多人场景教师识别	含多人的课堂录像	正确定位并追踪主讲教师	通过
特征提取	噪声环境音频处理	SNR=10dB 低信噪比录音	成功提取音频特征	通过
风格识别	标准样本分类	测试集已标注样本	风格标签与标注一致	通过
风格识别	混合风格识别	多风格混合的课堂片段	输出合理的概率分布	通过
画像生成	雷达图生成	7 维风格评分向量	正确渲染交互式雷达图	通过
画像生成	情绪曲线生成	时序情感标签序列	折线图按时间顺序渲染	通过
画像生成	关键词云图生成	课程全文转写文本	生成词频正确的云图	通过
追踪分析	风格稳定性计算	同一教师 10 节课记录	标准差计算正确	通过
追踪分析	SMI 相似度计算	当前评分与参考模板	SMI 值在 [0,1] 范围内	通过
SHAP 分析	特征归因计算	单次预测特征向量	返回 70 维 SHAP 值	通过
典型片段	片段自动提取	片段置信度列表	每类风格提取 Top-3 片段	通过
批量分析	多任务并行提交	5 节课同时提交分析	全部任务成功完成	通过
任务管理	任务状态查询	进行中的任务 ID	返回当前处理进度	通过
任务管理	失败任务重试	模拟服务异常后重试	自动重试后成功完成	通过

表 4.3　性能测试结果

测试场景	测试指标	目标值	实测值
单片段推理（已提取特征）	SHAPE 模型推理时间 [a]	≤1.5 秒	1.34 秒
整课完整流水线（45 分钟）	全流程总处理时间 [b]	≤75 分钟	约 62 分钟
缓存命中分析	重复分析响应时间	≤1 分钟	约 45 秒
并发用户（50 人）	系统响应时间	≤3 秒	2.1 秒
批量分析（35 节课）	总处理时间	≤90 分钟	58 分钟
SHAP 归因计算	单样本耗时	≤200 毫秒	120 毫秒

线（YOLOv8+DeepSORT+MediaPipe+ST-GCN）耗时 0.82 秒、音频流水线耗时 0.37 秒、SHAPE 推理耗时 0.15 秒，三路并行执行，总时间由最慢的视频流水线决定。

[b] 整课完整流水线时间说明：约 62 分钟包含视频解码与帧提取、Whisper Large-v3 全文语音识别（约 45 分钟课堂音频的 ASR 转写约耗时 20 分钟）、语义分段、所有片段特征提取（约 150 个片段顺序处理）以及 SHAPE 推理，是从原始视频输入到风格画像输出的完整流水线时间。各项性能指标均达到设计目标。

（二）安全性测试

系统进行了以下安全性验证：在接口访问控制方面，所有 API 接口均实现 JWT（JSON Web Token）身份认证，未授权请求返回 401 状态码，测试中连续尝试 50 次未授权访问均被正确拒绝；在文件上传安全方面，系统对上传文件进行 MIME 类型校验和文件头魔数验证，测试中上传的伪造扩展名可执行文件（将.exe 改名为.mp4）均被正确识别并拒绝；在 SQL 注入防护方面，系统通过 SQLAlchemy 参数化查询处理所有数据库操作，测试中构造的 SQL 注入载荷均未引发异常执行；在视频数据安全方面，MinIO 存储桶配置为私有访问策略，视频文件通过有效期为 15 分钟的预签名 URL 按需提供，有效防止敏感录像的未授权访问。

4.5.4　系统试运行

为验证系统在真实教育场景中的可用性，本研究对自建数据集的 55 节课录像（涵盖 7 位教师，总时长约 41 小时）进行了系统试运行，评估系统在完整使用流程中的稳定性与实用效果。

试运行期间，系统累计处理视频约 221GB，生成分析报告 55 份，完成 SHAP 归因计算 55 次，累计运行时间超过 60 小时，未发生系统崩溃或数据丢失事件。对于因网络波动导致上传中断的 3 次事件，断点续传功能均成功恢复，用户体验未受明显影响。任务处理阶段发生 2 次 Celery 工作进程异常退出，自动重试机制在 3 分钟内完成了任务的恢复执行，用户端无需手动干预。

在分析结果的可用性方面，参与试运行的 7 位教师在查看自己的风格画像报告后，均能通过雷达图和典型片段回放对系统的识别结果形成合理认知。其中 4 位教师表示风格雷达图与其自我认知” 基本一致”，2 位教师表示” 部分一致”，1 位教

师对某一维度的识别结果存有疑问，经查阅对应的 SHAP 特征详情后，认为该判断” 有一定道理但与自己预期不同”。整体来看，系统的可解释性设计有效降低了教师对结果的质疑程度，多维度的可视化输出也得到了教师的积极反馈。

4.6　本章小结

本章在第三章算法研究成果的基础上，完成了教师风格画像分析系统的设计与实现，将 SHAPE 多模态教师风格识别模型转化为面向教育场景的实用系统。

在**系统设计**方面，系统遵循模块化、可解释性和高性能三项核心原则，采用五层架构（数据管理层、特征提取层、模型推理层、应用服务层、用户交互层）组织各功能模块，通过微服务设计实现各层的独立部署与弹性扩展。系统支持单机部署（校内试点）和分布式部署（区域推广）两种模式，适应不同规模的应用场景。

在**技术实现**方面，前端采用 Vue 3 + ECharts 5.4 实现响应式交互界面与多维度可视化；后端采用 Flask + Celery + RabbitMQ 实现 RESTful API 服务与异步任务调度；模型推理采用 PyTorch + TensorRT 实现 GPU 加速，单片段处理时间 1.34 秒；数据存储采用 MySQL + Redis + MinIO 的分层架构，通过三级缓存将重复分析的响应时间降至 45 秒以内；系统通过 Docker Compose 实现容器化一键部署，降低运维门槛。

在**功能实现**方面，系统提供多模态特征提取、风格分类推理、可解释性分析（SHAP）、风格画像可视化（雷达图、行为柱状图、情绪曲线、关键词云、典型片段）、风格相似度评估（SMI）、风格稳定性追踪六大核心功能，覆盖从视频上传到风格报告生成的完整使用流程。

在**测试验证**方面，系统通过了 18 项功能测试（全部通过）、6 项性能测试（各项指标均达到设计目标）和多项安全性测试，并在 55 节课录像的试运行中保持稳定运行，未发生系统级故障。试运行中参与教师的反馈表明，系统的可解释性设计有效提升了画像结果的可信度与用户接受度，验证了将 SHAPE 算法工程化落地的可行性。

第五章　总结与展望

5.1　工作总结

本文主要介绍了基于多模态深度学习的教师教学风格画像分析系统的设计与实现，旨在解决传统课堂评价方法主观性强、反馈滞后、覆盖面窄等问题。

第二章对相关技术背景进行了系统综述。首先梳理了教学风格的理论分类体系，从 Flanders 互动分析到 Grasha 五维度框架，阐明了现有教学风格研究从主观评定向数据驱动转型的趋势。然后回顾了视频行为识别、语音情感分析、文本语义理解和多模态融合等支撑技术的发展脉络，为后续方法设计提供了理论依据。

第三章提出了 SHAPE（Semantic Hierarchical Attention Profiling Engine）多模态教师风格识别框架，这是本文的核心研究内容。首先提出了语义驱动的课堂分段策略，以 H-DAR 识别的教学意图边界替代固定时间窗口，消融实验（配置 B）证实该策略将风格识别准确率提升 7.6 个百分点（$p < 0.05$）。然后分别设计了三条模态特征提取流水线：视频模态通过 DeepSORT 身份追踪与 ST-GCN 时空图卷积网络提取 20 维骨骼动作特征；音频模态通过 Wav2Vec 2.0 自监督表征与情感分类器提取 15 维声学与情感特征；文本模态通过 BERT 结合层次化对话行为识别（H-DAR）提取 35 维教学意图特征。接着提出了双向差分跨模态注意力（BD-CMA）融合网络，将 BiXT 的 3 对双向共享相似度计算与 DiffAttention 的差分降噪机制融合，在降低计算冗余的同时抑制注意力噪声，配合 VMRNN 时序建模（$O(N)$ 线性复杂度）和注意力池化，完成 7 类教学风格的分类。最后设计了基于 SHAP 特征归因与跨模态注意力权重的可解释性分析框架，揭示了不同教学风格的模态依赖规律。在自建的 209 样本教师风格数据集上，通过多模态融合对比实验和消融实验对上述方法进行了系统评估。

第四章在 SHAPE 算法研究的基础上，完成了教师风格画像分析系统的工程实现。系统采用五层架构（数据管理层、特征提取层、模型推理层、应用服务层、用户交互层），前端基于 Vue 3 与 ECharts 5.4 构建，后端采用 Flask 与 Celery 异步任务框架，模型推理层通过 TensorRT 加速实现单片段 1.34 秒的端到端处理。系统提

供风格雷达图、行为柱状图、语音情绪曲线、关键词云图、典型片段回放等多维度可视化画像，并支持风格相似度评估（SMI）和跨课追踪分析。经功能测试、性能测试和 55 节课的试运行验证，系统运行稳定，各项指标均满足设计目标。

5.2 未来展望

尽管本研究在多模态教师风格识别与系统实现方面取得了一定成果，但仍存在若干有待改进之处，需要在后续研究中进一步深化。

在数据与模型层面，当前自建数据集规模为 209 个样本，数据来源主要集中于中学数学课堂，模型的跨学科、跨学段泛化能力尚未得到充分验证。未来需要扩充数据集规模，覆盖语文、物理、英语等多学科和小学、高中、大学等多学段的课堂录像，并建立更规范的多轮专家标注机制以进一步提升标注一致性。在模型性能方面，当前单片段处理时间为 1.34 秒，无法支持真正的实时课堂分析场景；可通过知识蒸馏和 INT8 量化压缩对 SHAPE 模型进行轻量化，将推理延迟降至 0.5 秒以内，并探索在录播终端的边缘部署方案。此外，当前模型假设三路模态信号均可用，对音频缺失或视频遮挡等情形的鲁棒性有待研究，未来可引入基于注意力门控的缺失模态补偿机制加以改善。

在应用与伦理层面，课堂视频涉及师生肖像权等隐私问题，需要在系统部署前建立完善的数据脱敏与访问控制机制。联邦学习框架可在不将原始视频传输出校园的前提下实现跨校模型协作训练，是兼顾隐私保护与模型泛化的重要方向。在功能扩展方面，引入眼动追踪、生理信号等新型模态数据，构建涵盖教师与学生双主体的课堂交互分析模型，有望进一步揭示教学风格与学生学习效果之间的内在关联，为教育研究提供更丰富的数据支撑。

参考文献

[1] N. Flanders. (1963). Intent, Action and Feedback: A Preparation for Teaching. *Journal of Teacher Education*, 14(3). DOI: 10.1177/002248716301400305.

[2] Grasha, A. F. (1996). *Teaching With Style: A Practical Guide to Enhancing Learning by Understanding Teaching and Learning Styles*. Pittsburgh: Alliance Publishers.

[3] Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) Manual, K-3*. Baltimore: Paul H. Brookes Publishing.

[4] 顾小清, 王炜. 信息技术整合课堂的互动分析系统研究 [J]. 中国电化教育, 2007(12): 48–53.

[5] 胡小勇, 眭慧, 陈莹, 穆肃. 多场景融合的教师数字画像：模式建构与应用方法 [J]. 中国远程教育, 2024, 44(4).

[6] Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up End-to-end Speech Recognition. *arXiv:1412.5567*.

[7] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems (NeurIPS)*, 33.

[8] Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML)*.

[9] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781*.

[10] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[11] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

[12] Simonyan, K., & Zisserman, A. (2014). Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems (NeurIPS)*, 27.

[13] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[14] Carreira, J., & Zisserman, A. (2017). Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast Networks for Video Recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[16] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C. L., Yong, M. G., Lee, J., Chang, W. T., Hua, W., Georg, M., & Grundmann, M. (2019). MediaPipe: A Framework for Perceiving and Processing Reality. *Third Workshop on Computer Vision for AR/VR at IEEE CVPR*.

[17] Yan, S., Xiong, Y., & Lin, D. (2018). Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby,

N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*.

[19] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., & Hu, H. (2022). Video Swin Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Qu, A., Wen, Y., & Zhang, J. (2025). ClassMind: Scaling Classroom Observation and Instructional Feedback with Multimodal AI. *arXiv:2509.18020*.

[21] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

[22] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016). Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *European Conference on Computer Vision (ECCV)*.

[23] Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLOv8 (Version 8.0.0) [Computer software]. Zenodo. DOI: 10.5281/zenodo.7347926.

[24] N. Wojke, Alex Bewley, D. Paulus. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. *2017 IEEE International Conference on Image Processing (ICIP)*. DOI: 10.1109/ICIP.2017.8296962.

[25] Ye, T., Dong, L., Xia, Y., Sun, Y., Zhu, Y., Huang, G., & Wei, F. (2024). Differential Transformer. *arXiv:2410.05258*.

[26] Tang, Y., Peng, P., Wen, Q., Zhou, T., Sun, L., & Jin, R. (2024). VMRNN: Integrating State Space Models and LSTMs for Efficient and Effective Spatiotemporal Forecasting. *arXiv:2405.16773*.

[27] Gu, A., & Dao, T. (2023). Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv:2312.00752*.

[28] H. H. Anderson. (1939). The Measurement of Domination and of Socially Integrative Behavior in Teachers' Contacts with Children. *Child Development*, 10(2). DOI: 10.1111/J.1467-8624.1939.TB04615.X.

[29] Zhou, Suping, Jia, Jia, Yin, Yufeng, Li, Xiang, Yao, Yang, Zhang, Ying, et al. (2019). Understanding the Teaching Styles by an Attention Based Multi-task Cross-media Dimensional Modeling. *Proceedings of the 27th ACM International Conference on Multimedia*. DOI: 10.1145/3343031.3351059.

[30] J. D. Guerrero-Sosa, Francisco P. Romero, V. Menéndez-Domínguez, J. Serrano-Guerrero, Andrés Montoro-Montarroso, J. A. Olivas. (2025). A Comprehensive Review of Multimodal Analysis in Education. *Applied Sciences*. DOI: 10.3390/app15115896.

[31] Ozan Raşit Yürüm. (2025). Technology-Enhanced Multimodal Learning Analytics in Higher Education: A Systematic Literature Review. *IEEE Access*. DOI: 10.1109/ACCESS.2025.3572467.

[32] 胡小勇, 林祥耀. 精准教研视域下的教师画像研究 [J]. 电化教育研究, 2019, 40(7): 21–27.

[33] 胡小勇, 孙硕, 穆肃. 基于画像技术的教师研修路径智能推荐研究 [J]. 电化教育研究, 2024(2). DOI: 10.13811/j.cnki.eer.2024.02.015

[34] 陈鑫, 胡东芳. 教师画像的前沿探讨：定义、概念框架与研究边界 [J]. 教师教育学报, 2021, 8(6): 1–8.

[35] 柏宏权, 朱俊. 小学人工智能教师画像构建研究 [J]. 电化教育研究, 2024(7). DOI: 10.13811/j.cnki.eer.2024.07.013

[36] 黄荣怀, 陈庚, 张进宝, 陈桄, 汪燕. 面向 K12 教师的智能教育素养框架构建 [J]. 开放教育研究, 2021, 27(4).

[37] 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).

[38] Qiaoyue Zhao. (2024). Design of Teacher Portrait System Based on Knowledge Graph. *2024 IEEE 2nd International Conference on Image Processing and Computer Applications (ICIPCA)*. DOI: 10.1109/ICIPCA61593.2024.10708989.

[39] Xiaoyong Hu, Hui Sui, Xingyu Geng, Li Zhao. (2024). Constructing a Teacher Portrait for the Artificial Intelligence Age Based on the Micro Ecological System Theory: A Systematic Review. *Education and Information Technologies*. DOI: 10.1007/s10639-024-12513-5.

[40] Lee, Unggi, Jeong, Yeil, Koh, Junbo, Byun, Gyuri, Lee, Yunseo, Lee, Hyunwoong, et al. (2024). I See You: Teacher Analytics with GPT-4 Vision-Powered Observational Assessment. *Smart Learning Environments*. DOI: 10.1186/s40561-024-00335-4.

[41] Conrad Borchers, Yeyu Wang, Shamya Karumbaiah, Muhammad Ashiq, D. Shaffer, Vincent Aleven. (2023). Revealing Networks: Understanding Effective Teacher Practices in AI-Supported Classrooms Using Transmodal Ordered Network Analysis. *Proceedings of the 14th Learning Analytics and Knowledge Conference*. DOI: 10.1145/3636555.3636892.

[42] Cong Wu, Xiaojun Wu, Tianyang Xu, Zhongwei Shen, J. Kittler. (2024). Motion Complement and Temporal Multifocusing for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. DOI: 10.1109/TCSVT.2023.3236430.

[43] Shaojie Zhang, Jianqin Yin, Yonghao Dang, Jiajun Fu. (2023). SiT-MLP: A Simple MLP With Point-Wise Topology Feature Learning for Skeleton-Based Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*. DOI: 10.1109/TCSVT.2024.3386553.

[44] Li, Yuanzhong, Deng, Zhengjie, Liu, Meijun, He, Shuqian, Wang, Yizhen, Jiang, Wenjuan. (2022). A Method for Analyzing Teacher Behavior in Classroom Based

on the Long- and Short-Term Features of Pose Sequences. *2022 9th International Conference on Digital Home (ICDH)*. DOI: 10.1109/icdh57206.2022.00043.

[45] Cai, Ting, Xiong, Yu, He, Chengyang, Wu, Chao, Cai, Linqin. (2025). Classroom Teacher Behavior Analysis: The TBU Dataset and Performance Evaluation. *Computer Vision and Image Understanding*. DOI: 10.1016/j.cviu.2025.104376.

[46] Zhu, Wenqi, Yang, Zhijun. (2024). Csb-yolo: a Rapid and Efficient Real-time Algorithm for Classroom Student Behavior Detection. *Journal of Real-Time Image Processing*. DOI: 10.1007/s11554-024-01515-8.

[47] Óscar Cánovas Reverte, Félix J. García Clemente, Federico Pardo. (2023). AI-driven Teacher Analytics: Informative Insights on Classroom Activities. *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*. DOI: 10.1109/TALE56641.2023.10398309.

[48] Ziyi Liu, Li Yang, Sannyuya Liu. (2023). A Teacher and Student Facial Expression Recognition Model Based on Classroom Teaching Videos. *2023 International Conference on Intelligent Education and Intelligent Research (IEIR)*. DOI: 10.1109/IEIR59294.2023.10391209.

[49] Zhao, Lanfei, Lin, Zixiang, Sun, Ruiyang, Wang, Aili. (2024). A Review of State-of-the-Art Methodologies and Applications in Action Recognition. *Electronics*. DOI: 10.3390/electronics13234733.

[50] Fatemeh Shafizadegan, A. Naghsh-Nilchi, Elham Shabaninia. (2024). Multimodal Vision-Based Human Action Recognition Using Deep Learning: A Review. *Artificial Intelligence Review*. DOI: 10.1007/s10462-024-10730-5.

[51] Schiappa, Madeline C., Rawat, Yogesh S., Shah, Mubarak. (2023). Self-Supervised Learning for Videos: A Survey. *ACM Computing Surveys*. DOI: 10.1145/3577925.

[52] Olfa Saket, A. B. Aicha, H. Fathallah. (2025). Deep Learning Applied for Abnormal Human Behavior Recognition in Video Surveillance Systems: A Systematic Review. *Applied Intelligence*. DOI: 10.1007/s10489-025-06797-4.

[53] Chen-Lin Zhang, Jianxin Wu, Yin Li. (2022). ActionFormer: Localizing Moments of Actions with Transformers. *European Conference on Computer Vision (ECCV)*. DOI: 10.1007/978-3-031-19772-7_29.

[54] Mu, F., Mo, S., Wang, G., & Li, Y. (2022). Where a Strong Backbone Meets Strong Features – ActionFormer for Ego4D Moment Queries Challenge. *arXiv:2211.09074*.

[55] Ding Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, Dacheng Tao. (2023). TriDet: Temporal Action Detection with Relative Boundary Modeling. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR52729.2023.01808.

[56] Fangzhou Mu, Sicheng Mo, Yin Li. (2024). SnAG: Scalable and Accurate Video Grounding. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR52733.2024.01791.

[57] M. Rafique, Faheem Khaskheli, Malik Tahir Hassan, Sheraz Naseer, M. Jeon. (2022). Employing Automatic Content Recognition for Teaching Methodology Analysis in Classroom Videos. *PLoS ONE*. DOI: 10.1371/journal.pone.0263448.

[58] J. Riordan, Lynn Revell, B. Bowie, Sabina Hulbert, M. Woolley, Caroline Thomas. (2024). Multimodal Classroom Interaction Analysis Using Video-Based Methods of the Pedagogical Tactic of (un)grouping. *Pedagogies: An International Journal*. DOI: 10.1080/1554480X.2024.2313978.

[59] Liang Jie, Xiaoyan Zhao, Zhaohui Zhang. (2020). Speech Emotion Recognition of Teachers in Classroom Teaching. *2020 Chinese Control And Decision Conference (CCDC)*. DOI: 10.1109/CCDC49329.2020.9164823.

[60] Donnelly, P., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., & D'Mello, S. K. (2016). Automatic Teacher Modeling from Live Classroom Audio. *Proceedings of the 2016 Conference on User Modeling, Adaptation and Personalization (UMAP)*.

[61] Federico Pardo, Óscar Cánovas, Félix J. García Clemente. (2025). Audio Features in Education: A Systematic Review of Computational Applications and Research Gaps. *Applied Sciences*. DOI: 10.3390/app15126911.

[62] Zhang, Xilin, Wang, Jiaqi, Wan, Zhenhong, Luo, Zuying. (2021). Classification of Classroom Teachers' Speech Intention Based on Deep Learning. *2021 International Conference on Computer Engineering and Application (ICCEA)*. DOI: 10.1109/iccea53728.2021.00053.

[63] Qing Han, L. Luo, Zhong Sun. (2020). Research on Teachers' Behavior in the Class Recognition on Based on Text Classification Technology. *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*. DOI: 10.1109/ICEIEC49280.2020.9152304.

[64] Óscar Sapena, E. Onaindía. (2022). Multimodal Classification of Teaching Activities from University Lecture Recordings. *Applied Sciences*, 12(9). DOI: 10.3390/app12094785.

[65] Ahmad, Kashif, Qadir, Junaid, Al-Fuqaha, Ala, Iqbal, Waleed, El-Hassan, Ammar, Benhaddou, Driss, et al. (2020). Data-Driven Artificial Intelligence in Education: A Comprehensive Review. *OSF Preprints*. DOI: 10.35542/osf.io/zvu2n.

[66] S. Howard, Jie Yang, Jun Ma, C. Ritz, Jiahong Zhao, K. Wynne. (2018). Using Data Mining and Machine Learning Approaches to Observe Technology-Enhanced Learning. *2018 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*. DOI: 10.1109/TALE.2018.8615443.

[67] Housen, H. T. (2020). Lecture2Notes: Summarizing Lecture Videos by Classifying Slides and Analyzing Text. *arXiv preprint*.

[68] Qifang Liu, Wuying Deng. (2022). Animation User Value Portrait Based on RFM Model under Big Data. *Mathematical Problems in Engineering*. DOI: 10.1155/2022/8246540.