

从信号到意义

课堂协作与投入度的多模态学习分析

涂蓓
吉林艺术学院

摘要 本文系统综述多模态学习分析（MMLA）在课堂协作与学习投入度研究中的最新进展。研究聚焦视觉、听觉、生理与交互数据的融合机制，比较早期、晚期及基于注意力的混合建模方法，提出评估鲁棒性、脱模态泛化与时序对齐的可复现协议。通过引入认知负荷理论与文化资本视角，本文揭示了学习状态测量与教育公平之间的内在张力，并讨论了 PIPL GDPR 双栈治理框架、在地化设计原则及多模态大语言模型（MLLM）的零样本应用前景。研究旨在为课堂学习状态的智能评估提供兼具技术可行性与伦理适切性的整体范式。

关键词 多模态学习分析；课堂协作；学习投入度；跨模态注意力；伦理治理

DOI <https://doi.org/10.6914/aiescn.010203> **文章编号** 3079-7918.2025.0102.30-46

引言：从单模态到多模态——学习分析的范式革新

（一）传统学习分析的局限性

学习分析（Learning Analytics, LA）被权威地定义为“测量、收集、分析和报告关于学习者及其学习情境的数据，以理解和优化学习及其发生的环境”的过程。^① 在过去十余年的发展中，该领域的研究主要集中于分析学习管理系统（Learning Management System, LMS）、大规模开放在线课程（MOOCs）等数字化学习平台所产生的海量数据。^② 这些研究普遍依赖于所谓的“单维度”（unidimensional）数据，即学习者与系统交互时留下的数字痕迹，例如页面浏览、视频观看时长、论坛发帖和作业提交等点击流数据。

© The Author(s) 2025. **Article History** Received: 14 July 2025; Accepted: 19 August 2025; Published: 30 September 2025. 人工智能教育研究 (Réngōng Zhìnéng Jiàoyù Yánjiū) *Artificial Intelligence Education Studies*, ISSN 3079-7918 (Print), ISSN 3079-7926 (Online). Hosted by **Artificial Intelligence Education Studies Press** (Certificate No.70125648-007), and published by **Creative Publishing Co., Limited** (CPCL®) in Hong Kong. Email: cpcl@cpcl.hk, wtoom@gmail.com, <https://aies.press>, <https://cpcl.hk>. Repository: <https://pdf.ssci.cc/index.html>. This is an open-access article distributed under the terms of the Creative Commons Attribution-NoDerivatives 4.0 International License, which permits use, sharing, and redistribution in any medium or format, provided the original author(s) and source are credited, but does not allow modifications or adaptations of the work.

^①George Siemens and Dragan Gašević, “Guest Editorial-Learning and Knowledge Analytics”, 《Educational Technology & Society》, 2012, 15(3), pp. 1-2.

^②参见 Dragan Gašević, Shane Dawson, and George Siemens, “Let’s not forget: Learning analytics are about learning”, 《TechTrends》, 2015, 59, pp. 64-71.

尽管基于点击流的分析在预测学生辍学风险、评估学习行为模式等方面取得了一定的成功，但其固有的局限性也随着研究的深入而日益凸显。其核心问题在于，这类数据在很大程度上忽略了学习者所处的丰富情境信息（contextual information），导致对学习过程的理解存在“情境盲视”（contextual blindness）。真实的学习过程，尤其是发生在物理课堂中的学习，是一个复杂、动态且充满社会性互动的过程。单一的点击流数据难以捕捉师生之间、生生之间的非言语交流（如面部表情、手势、身体姿态）、口头话语的细微差异（如语调、语速）、学习者内在的认知负荷与情绪波动，也无法反映物理空间中的协作动态。为了实现对学习过程的精准理解与有效干预，教育数据科学领域迫切需要超越单一的数字痕迹，转向能够采集并分析反映学习全貌的多维度、多来源数据。

（二）多模态学习分析（MMLA）的兴起与价值

为应对传统学习分析的挑战，多模态学习分析（Multimodal Learning Analytics, MMLA）应运而生。MMLA 被定义为一个新兴的学习分析前沿领域，它通过整合来自不同来源的多种数据模态（如视频、音频、生理信号、数字交互日志等），以实现对学习过程的“全息”（holistic）理解。^① MMLA 的核心价值在于，它不再依赖单一数据源，而是通过数据三角测量（triangulation）的方式，将语音、视线、姿态和生理信号等多种数据类型进行整合，从而为复杂的学习过程提供更丰富、更全面的表征。^②

这种范式革新不仅仅是数据量的增加或数据种类的叠加，更代表了一种从分析学习的“结果代理”到观察学习“动态过程”的质性飞跃。传统 LA 捕捉的是离散的、已完成的动作（如一次点击、一次提交），这些是学习过程的间接代理指标。而 MMLA 捕捉的是连续的数据流，如随时间变化的视线坐标、不断波动的声学特征、以及实时变化的脑电波段功率。这种面向过程的数据使得研究者能够对学习的核心构念，如认知负荷、注意力波动、情绪唤醒等进行建模，从而揭示了学习者行为与学习结果之间的潜在机制或中介过程。正是这种对学习过程的深度洞察，使得构建真正意义上的“个性化和自适应学习环境”成为可能，因为教育干预可以基于学习者的实时状态，而非仅仅是他们过往的行为记录。

（三）MMLA 在 LAK 与 SoLAR 社群中的历史与理论奠基

MMLA 的学术谱系与学习分析领域的核心学术社群——学习分析研究学会（Society for Learning Analytics Research, SoLAR）及其旗舰会议——国际学习分析与知识大会（International Conference on Learning Analytics & Knowledge, LAK）的发展紧密相连。^③ 这些平台为 MMLA 的理论构建、方法论探讨和社群形成提供了关键的孵化土壤。SoLAR 通过其官方渠道，不断更新和探讨学习分析的定义与边界，以适应技术和理论的演进。^④

SoLAR 通过出版《学习分析期刊》（Journal of Learning Analytics）和《学习分析手册》（Handbook of Learning Analytics）等权威刊物，系统性地确立了 MMLA 作为 LA 领域一个重要分支的学术地位。^⑤ 早期的《学习分析手册》章节便已明确提出，MMLA 旨在通过捕获、融合和分析互补的学习痕迹来源，以获得对学习过程更稳健、更少不确定性的理解，其数据模态不仅包括传统的日志文件，还涵盖了手势、视线、言语等更自然的人类信号。^⑥ 这为 MMLA 的研究议程设定了基调。

^①J.D.T. Guerrero-Sosa et al., “A Comprehensive Review of Multimodal Analysis in Education”, 《Applied Sciences》, 2025, 15(11), 5896.

^②Chango, W. et al., “Multimodal Data Fusion in Learning Analytics: A Systematic Review”, 《Sensors》, 2020, 20(23), 6866.

^③George Siemens, “Learning Analytics: A Foundation for a Field”, in Charles Lang, George Siemens, Alyssa Wise, and Dragan Gašević eds., 《The Handbook of Learning Analytics》, 1st ed., Edmonton: Society for Learning Analytics Research, 2017, pp. 8-16.

^④Society for Learning Analytics Research, “SoLAR’s Learning Analytics Definition Taskforce releases report”, 2025 年 6 月 18 日, <https://www.solaresearch.org/2025/06/solars-learning-analytics-definition-taskforce-releases-report/>, 2025 年 10 月 3 日。

^⑤Charles Lang, George Siemens, Alyssa Wise, Dragan Gašević, and Agathe Merceron eds., 《Handbook of Learning Analytics》, 2nd ed., New York: Society for Learning Analytics Research, 2022.

^⑥Xavier Ochoa, “Multimodal Learning Analytics: Rationale, Process, Examples, and Direction”, in Charles Lang, George Siemens, Alyssa Wise, Dragan Gašević, and Agathe Merceron eds., 《Handbook of Learning Analytics》, 2nd ed., New York: Society for Learning Analytics Research, 2022, pp. 54-65.

LAK 会议则成为 MMLA 研究成果发布和思想碰撞的主要阵地。历届 LAK 会议的论文集和主题演讲记录了 MMLA 从概念走向实践的演进轨迹。例如,早期的会议就已经开始探讨如何利用多模态数据来理解具身互动学习环境。随着技术的成熟,SoLAR 内部成立了专门的兴趣小组,如“跨空间多模态学习分析特别兴趣小组”(Cross-MMLA SIG),旨在推动研究社群共同应对在复杂教育情境中(涉及多种模态、多个参与者和多个学习空间)理解数据的挑战。^①这一系列举措,包括举办年度工作坊和学术专刊,极大地促进了一个充满活力的研究社群的形成,不断推动着 MMLA 研究的前沿。这些奠基性的工作为本综述提供了坚实的史观基础,使其能够将当前的技术进展置于一个清晰的学术发展脉络之中。

(四) 研究问题与论文结构

尽管 MMLA 前景广阔,但其在真实课堂环境中的应用仍处于探索阶段,面临着理论、技术和伦理等多方面的挑战。本综述旨在提出并遵循一个独创的“方法—证据—规范”三维分析框架,系统性地梳理和探讨基于 MMLA 的课堂协作与投入度自动评估问题。这一框架本身构成本文的核心理论贡献,它提供了一个整合性的视角,用以评估和指导 MMLA 研究的发展。本文力图回答以下核心问题:

方法(Method):如何构建一个从理论操作化到数据融合建模的可复现技术框架,以有效评估课堂中动态、复杂的协作与投入度?其中涉及哪些关键的多模态数据通道、特征提取技术,以及如何从历史基线模型(如 LSTM)演进到当前最优范式(如 Transformer)?

证据(Evidence):如何将“多模态优于单模态”的抽象论断转化为可检验的实证协议?即,如何通过量化指标(如噪声鲁棒性、脱模态泛化能力)为 MMLA 的核心价值——“消歧”——提供确凿证据?

规范(Norm):在真实课堂中应用 MMLA 面临哪些核心的技术挑战、伦理困境与治理难题?如何构建一个在操作层面符合国际数据保护法规(如《个人信息保护法》PIPL 与《通用数据保护条例》GDPR)并兼顾教育情境特殊性的治理框架?

前沿(Frontier):以多模态大语言模型(Multimodal Large Language Models, MLLMs)为代表的新兴技术,如何重塑 MMLA 的研究范式?如何将其从一个概念性的“机遇”转化为一个可复现、可评估、风险可控的最小闭环应用范式?

为回答上述问题,本文结构安排如下:第二部分构建 MMLA 的理论框架,将核心构念操作化,并明确提出“方法—证据—规范”作为本文的独创分析框架,同时引入认知负荷理论与文化资本理论作为跨学科透镜;第三部分详细阐述面向课堂分析的各模态特征工程技术;第四部分聚焦多模态数据融合策略与深度学习建模方法;第五部分为 MMLA 的核心价值“消歧”建立一个标准化的评测协议;第六部分补充一个基于开源数据集的实证案例,以展示评测协议的可操作性;第七部分将前沿的 MLLMs 范式具体化为可操作的应用框架;第八部分对 MMLA 在实践中面临的挑战和伦理问题进行批判性反思,并提出在地化解决方案与双栈合规治理框架,同时进行国际比较;第九部分凝练本文的核心贡献;最后对全文进行总结。

一、理论框架:将学习构念操作化以进行多模态分析

(一) 学习投入度的三维模型:一个奠基性的理论视角

在将 MMLA 应用于课堂分析之前,必须首先为所关注的核心构念建立一个坚实的理论基础。对于“学习投入度”(Learning Engagement)这一核心构念,学术界广泛采用的奠基性框架是由 Fredricks, Blumenfeld, & Paris (2004) 提出的三维模型。^②该模型将投入度解构为三个相互关联但又各自独立的维度:

^①Society for Learning Analytics Research, “Multimodal Learning Analytics Across Spaces (SOLAR CrossMMLA SIG)”, <https://www.solaresearch.org/community/sigs/crossmmla-sig/>, 2025 年 10 月 3 日。

^②J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, “School Engagement: Potential of the Concept, State of the Evidence”, 《Review of Educational Research》, 2004, 74(1), pp. 59-109.

行为投入、情感投入和认知投入。这一框架之所以重要，因为它为 MMLA 的数据采集和特征提取提供了清晰的理论指引，即将抽象的心理状态与可观察、可测量的多模态信号联系起来。

行为投入 (Behavioral Engagement)：指学习者在学习活动中的参与程度和努力。在 MMLA 的语境下，它可以通过一系列可观察的指标进行操作化。例如，通过视线追踪分析学生注视教师或教学材料的时长，可以量化其注意力；通过身体姿态分析（如是否前倾、是否记笔记）来判断其专注度；通过音频分析中的语音活动检测 (VAD) 来评估其课堂参与；或通过分析其在数字平台上的任务完成率和交互频率来评估其任务努力程度。

情感投入 (Affective Engagement)：指学习者在学习过程中伴随的情绪反应，如兴趣、愉悦、焦虑或厌倦。MMLA 主要通过非侵入式的方式来捕捉这些情感信号。例如，利用计算机视觉技术自动识别学生的面部表情（如微笑、皱眉），或通过分析语音的声学特征（如音高、语速、能量）来识别其语音情感，从而推断其情感状态。

认知投入 (Cognitive Engagement)：指学习者在完成学习任务时所投入的心理努力、使用的认知与元认知策略的深度。这是三个维度中最难直接测量的，也是 MMLA 展现其独特价值之处。MMLA 通过多种途径对其进行间接推断：一是通过自然语言处理技术分析学生的课堂话语或书面文本，评估其提问深度、论证复杂度和知识建构水平；二是通过生理传感器直接测量与心智努力密切相关的生理指标，最典型的是利用脑电图 (EEG) 信号来评估学生的认知负荷 (Cognitive Load)。

（二）协作学习的分析：超越个体，关注互动过程

对协作学习 (Collaborative Learning) 的评估同样是多维度的，其核心在于从关注协作的“结果”转向关注协作的“动态过程”和“质量”。这需要引用协作话语分析领域的教育学根基，这些研究强调了互动在知识共建中的核心作用。MMLA 技术为量化这些稍纵即逝的互动过程提供了前所未有的工具。

协作互动动态 (Interaction Dynamics)：指小组成员间的互动模式和结构。MMLA 可以通过音频分析量化每个成员的发言时长、发言频率和话轮转换 (turn-taking) 模式，从而评估参与的均衡度。同时，通过视觉分析，可以追踪小组成员间的视线同步性 (gaze synchronization)，将其作为联合注意 (joint attention) ——即小组成员是否关注同一对象的关键指标。

协作话语质量 (Discourse Quality)：指协作过程中语言交流的认知深度。通过对课堂语音进行自动语音识别 (Automatic Speech Recognition, ASR) 并转化为文本后，可以利用自然语言处理 (Natural Language Processing, NLP) 技术对话语内容进行分析，识别其中是否包含提问、解释、论证、质疑和共识建构等促进深度学习的关键词语行为。这些话语模式的识别，直接源于教育话语分析的理论传统。

（三）“多对多”映射难题与多模态消歧的理论必然性

对上述构念的操作化揭示了一个贯穿 MMLA 的核心复杂性：可观察的数据特征与潜在的心理状态之间存在着一种“多对多” (many-to-many) 的映射关系。例如，一个单一的特征，如“身体前倾”，可能既表示因高度投入而专注，也可能表示因困惑而努力辨识；反之，一个单一的状态，如“认知投入”，可能通过多种特征表现出来，如更复杂的话语结构、脑电 θ 波段功率的增加以及眨眼频率的降低。这种固有的模糊性意味着任何单一模态的分析都可能产生歧义。

这一现象不仅是 MMLA 面临的一个挑战，更是论证其核心方法论——多模态数据融合——必要性的根本理论基石。如果每个可观察特征都能唯一地对应一个内部状态，那么简单的单模态分析就已足够。然而，正是因为这种“多对多”的模糊性，使得单一信号的解读充满了不确定性。因此，数据融合的科学目标，并非简单地将数据“相加”以期获得微小的精度提升，其核心在于通过引入来自其他模态的上下文信息来“消歧” (disambiguation)。

这种对“消歧”的理论需求，为后续章节中讨论的先进融合模型（如跨模态注意力机制）提供了清晰的评价标准和设计目标。一个融合模型的优劣，很大程度上取决于它能否有效地解决这种由“多对多”映

射关系带来的不确定性。这构成了从理论问题到技术解决方案的坚实逻辑链条：

理论根源：人类行为与内在状态的映射关系是模糊的、多对多的。

问题表现：单一模态的信号解读具有歧义性。

方法论需求：需要一种机制，能够利用一个模态的上下文来澄清另一个模态的意义。

技术实现：这种需求在计算层面，最终由能够动态建模模态间条件依赖关系的跨模态注意力机制等先进模型来实现。

这种从理论难题到技术需求的演进逻辑，清晰地展示了为何简单的早期或晚期融合策略在处理复杂人类行为时存在局限，并凸显了动态、情境化融合方法的理论优越性。

（四）一个整合性的分析框架：作为独特贡献的“方法—证据—规范”

当前 MMLA 领域的综述文献，大多侧重于对技术方法或应用场景的梳理与盘点。^① 尽管这些工作极具价值，但它们往往将技术可行性、实证有效性与伦理社会影响作为分离的议题进行讨论，缺乏一个能够将三者内在联系起来的统一分析框架。为了弥补这一不足，本综述明确提出一个独创的“方法—证据—规范”（Method Evidence Norm）三维分析框架，并将其作为本文的核心理论贡献。

该框架主张，对任何 MMLA 系统的评估和设计，都必须同时在这三个维度上进行审视，三者缺一不可，且相互制约：

方法（Method）维度：关注技术层面的“如何做”，即系统的技术可行性与可复现性。它要求研究清晰地阐述其数据采集、特征提取、模型构建的全流程，确保其在科学上是严谨和可重复的。

证据（Evidence）维度：关注实证层面的“做什么”，即系统的经验有效性与核心价值。它要求研究超越简单的性能指标（如准确率），为 MMLA 的核心价值——“消歧”——提供可量化的、可检验的证据，证明多模态确实比单模态带来了质的提升。

规范（Norm）维度：关注伦理与社会层面的“为何做”与“为谁做”，即系统的伦理稳健性与社会适切性。它要求研究不仅要遵守数据保护法规，更要前瞻性地思考技术在真实教育情境中可能带来的偏见、不公以及对师生关系的潜在影响。

这一框架的独创性在于，它将三个原本分散的考量整合为一个相互关联的整体。一个在“方法”上先进的系统，如果不能在“证据”上证明其独特的消歧能力，或者在“规范”上存在伦理风险，那么它就不是一个值得推广的系统。反之，一个在“规范”上考虑周全的设计，也必须有可行的“方法”和可靠的“证据”来支撑。因此，“方法—证据—规范”框架为 MMLA 研究提供了一个更全面、更具批判性的评估标准，旨在推动该领域从单纯的技术驱动，转向技术、实证与价值协同驱动的成熟发展阶段。这一框架也为中国语境下的 MMLA 研究提供了独特的审视视角，确保技术发展与我国的教育现实、伦理法规紧密结合。

（五）跨学科透镜：认知负荷与文化资本的引入

为了进一步增强本综述的跨学科深度，我们在理论框架中引入了来自心理学和教育社会学的两个关键理论透镜，以丰富对 MMLA 数据和应用的阐释。

认知负荷理论（Cognitive Load Theory, CLT）：CLT 是一个解释大脑在学习过程中如何处理信息，特别是如何管理有限的工作记忆资源的教学设计理论。^② 它区分了三种认知负荷：由材料本身复杂性决定的内在负荷（intrinsic load），由教学呈现方式不当造成的外在负荷（extraneous load），以及用于构建心智图式以促进长期记忆的相关负荷（germane load）。^③ 将 CLT 与 MMLA 相结合，为“认知投入”的

^①参见 Hamza Ouhaichi, Daniel Spikol, and Bahtijar Vogel, “Research trends in multimodal learning analytics: A systematic mapping study”, 《Journal of Computer Assisted Learning》, 2023, 39(5), pp. 1325-1345.

^②John Sweller, Jeroen J. G. van Merriënboer, and Fred G. W. C. Paas, “Cognitive Architecture and Instructional Design”, 《Educational Psychology Review》, 1998, 10(3), pp. 251-296.

^③详见 The Decision Lab, “Cognitive Load Theory”, <https://thedecisionlab.com/reference-guide/psychology/cognitive-load-theory>, 2025 年 10 月 3 日。

测量提供了更精细的理论依据。MMLA 通过生理信号（如 EEG、瞳孔变化）直接测量总体的认知负荷，其理论目标不仅是识别高负荷状态，更是要区分这种负荷的来源。例如，一个理想的 MMLA 系统应能帮助教师判断，学生的高认知负荷究竟是源于任务本身的挑战性（高内在负荷，有益），还是源于教学材料呈现混乱（高外在负荷，有害）。这使得 MMLA 从一个简单的“状态监测器”转变为一个强大的“教学诊断工具”。^①

布迪厄的文化资本（Bourdieu's Cultural Capital）：法国社会学家皮埃尔·布迪厄提出的文化资本理论，为我们审视 MMLA 的社会公平性问题提供了深刻的批判性视角。^② 文化资本指个人拥有的非经济性社会资产，如语言风格、行为举止、文化知识等，它们能够影响个人在社会（尤其是教育系统）中的成功机会。^③ 布迪厄认为，教育系统往往会不自觉地偏爱和奖励那些拥有与主流文化相符的文化资本的学生，从而实现社会阶层的再生产。^④ 这一理论对 MMLA 提出了严峻的警示：如果一个 MMLA 系统的算法模型，是在特定（通常是主流）学生群体的行为数据上训练的，那么它可能会将该群体的行为模式（如特定的提问方式、互动节奏）视为“高投入”或“优良协作”的黄金标准。当这个模型被应用于具有不同文化资本（如不同方言、不同非言语交流习惯）的学生时，就可能发生系统性的误判，将差异解读为缺陷。这种算法层面的偏见，正是布迪厄所说的“符号暴力”（symbolic violence）在数字时代的体现，它以客观、中立的技术面貌，掩盖了对非主流文化的贬低与排斥。^⑤

将 CLT 与文化资本理论并置，揭示了 MMLA 设计中一个深刻的内在张力。从 CLT 的视角看，MMLA 的目标是追求教学效率的最大化，通过优化呈现方式来降低外在负荷。而从文化资本的视角看，这种对“最优”呈现方式的追求，本身就可能内嵌了文化偏见，其“优化”的可能只是针对拥有特定文化资本的学生群体。因此，一个负责任的 MMLA 系统设计，必须在这两种理论的张力之间寻求平衡：既要利用 CLT 的原理来提升认知效率，又要运用文化资本的透镜来不断进行自我审视和偏见审计，确保技术赋能的普惠性，而非加剧教育不公。

二、面向课堂分析的多模态特征工程与提取技术

特征工程是将原始的多模态数据转化为可供机器学习模型使用的结构化特征的过程，是决定 MMLA 系统性能的关键环节。然而，在这一技术过程背后，存在一个贯穿始终的根本性方法论张力：即在真实课堂（“in the wild”）中收集数据的生态效度（ecological validity）与确保模型所需信号质量（signal quality）之间的内在矛盾。MMLA 的最终目标是理解真实、复杂的学习情境，但这恰恰意味着数据采集环境是不可控和充满噪声的。本章将围绕这一核心张力，探讨视觉、听觉和生理三大主要通道的特征提取技术。

（一）视觉通道分析：洞察非言语行为

视觉通道是捕捉课堂非言语行为最直观、信息最丰富的模态。在理想的实验室条件下，高分辨率摄像头可以清晰地捕捉到学生的每一个细微表情和动作。

^①Mutlu Cukurova et al., “The role of learning theory in multimodal learning analytics”, 《British Journal of Educational Technology》, 2023, 54(5), pp. 1247-1266.

^②Pierre Bourdieu, “The Forms of Capital”, in John G. Richardson ed., 《Handbook of Theory and Research for the Sociology of Education》, New York: Greenwood Press, 1986, pp. 241-258.

^③参见 Saul McLeod, “Cultural Capital Theory of Pierre Bourdieu”, 《Simply Psychology》, 2023, <https://www.simplypsychology.org/cultural-capital-theory-of-pierre-bourdieu.html>, 2025 年 10 月 3 日。

^④Pierre Bourdieu and Jean-Claude Passeron, 《Reproduction in Education, Society and Culture》, London: Sage Publications, 1977.

^⑤参见 Gabe Fink, “The Theory of Cultural Capital in Higher Education and Its Influence on Equity”, Every Learner Everywhere, 2023, <https://www.everylearnereverywhere.org/blog/the-theory-of-cultural-capital-in-higher-education-and-its-influence-on-equity/>, 2025 年 10 月 3 日。

面部表情与情感识别：除了识别喜、怒、哀、惊等基本情绪外，更细粒度的分析方法是提取“面部动作单元”（Facial Action Units, AUs）。特定的 AUs 组合与学习过程中的复杂情感状态（如投入、困惑、挫败感）高度相关。

头部姿态与视线追踪：头部和眼睛的朝向是注意力的重要代理指标。在协作学习场景中，“视线同步性”（gaze synchronization）被认为是衡量“联合注意”（joint attention）和协作质量的有力指标。

身体姿态与动作分析：研究发现，身体姿态（如前倾、后仰）与学习者的投入状态相关。在协作任务中，小组成员间的“身体同步性”（body synchrony）往往预示着更高的协作融洽度。

然而，当这些技术被部署到真实课堂时，信号质量会受到严重挑战。多变的照明条件、学生频繁的身体移动和相互遮挡，都会对基于计算机视觉的算法的鲁棒性提出严峻考验。

（二）听觉通道分析：解码课堂话语互动

音频是分析课堂语言互动和情感氛围的核心模态。在安静的录音棚中，可以提取出高质量的话语和情感特征。

课堂话语结构分析：通过语音活动检测（VAD）和说话人日志（Speaker Diarization），可以计算出每个学生的总发言时长、发言频率、话轮转换模式等互动指标，直观反映学生的参与度。

自动语音识别与内容分析：将语音转化为文本（ASR）是理解话语认知内涵的关键。一旦获得转录文本，就可以利用 NLP 技术分析讨论的认知深度，评估其知识建构水平。

语音情感识别（SER）：通过分析语音信号的韵律特征（如音高、能量）和谱特征（如 MFCCs），可以识别说话者的情感状态，为评估学生的情感投入和协作氛围提供线索。

但在真实的课堂环境中，背景的嘈杂声、多人同时说话造成的串音、桌椅移动声等，都会严重污染音频信号，为 ASR 等技术带来了严峻的技术挑战，降低特征提取的准确性。特别是在多说话人场景下，需要采用先进的 ASR 域适配策略，例如，通过语音分离技术（如波束成形或深度学习掩码）来区分不同声源，结合基于语言模型的重打分机制，并通过融合多口音数据的自训练方法来提升模型对地方方言和不同口音的鲁棒性。

（三）生理通道分析：测量内在认知状态（机理修正）

生理信号为直接、客观地测量学习者内在的、不可见的认知与情感状态提供了独特的窗口。

脑电（EEG）与认知负荷（修正版）：脑电信号，特别是特定频段的功率变化，是评估认知负荷最直接和广泛研究的生理指标之一。与原稿中存在矛盾的表述不同，基于对大量神经科学文献的系统性回顾，认知负荷与脑电活动之间的关系可以被更精确地统一描述。在高认知负荷条件下，常见模式为额叶 θ 功率上升、顶叶 α 功率抑制（下降）；该模式亦与瞳孔扩大、EDA 基线提升等协同变化相一致。具体来说：

额叶 θ 波段（Frontal Theta, 4–8 Hz）功率上升：大量研究表明，额叶区域的 θ 波活动与工作记忆、认知控制和心智努力的增加密切相关。当学习者执行需要持续注意力和信息处理的复杂认知任务时，额叶 θ 功率的增强被认为是认知资源调动的一个可靠标志。

顶叶 α 波段（Parietal Alpha, 8–12 Hz）功率下降： α 波段的功率通常与大脑的“空闲”或抑制状态呈负相关。当需要进行外部信息处理或内部认知操作时，相关脑区的 α 功率会受到抑制，即发生所谓的“事件相关去同步化”（Event-Related Desynchronization, ERD）。特别是在顶叶区域， α 波的抑制被广泛认为是皮层激活和注意力资源分配的标志，因此认知负荷越高，顶叶 α 功率下降越显著。

对立证据与边界条件：然而，必须指出，这一模式并非绝对。其效应的方向和显著性受到多种因素的调节，包括任务类型（例如，一些创造性思维任务可能反而与 α 的增加相关）、个体差异（如认知能力、心理韧性等）、设备密度与分析方法（如频段的个体化定义）等。部分研究甚至报告了认知负荷与 θ 功率的负相关性，这凸显了在具体应用中谨慎解释 EEG 信号的必要性。

皮肤电（EDA）与情绪唤醒：皮肤电活动是衡量情绪唤醒度（emotional arousal）最常用的生理指标

之一。其缓慢变化的强直成分（SCL）反映总体唤醒水平，而对特定刺激的快速响应（SCR）则反映即时情绪反应。

生理信号测量的主要挑战在于其侵入性。佩戴脑电帽、生理手环等设备本身就可能对正常的教学活动构成干扰，让师生感到不适或不自然，从而影响其行为的真实性，导致所谓的“霍桑效应”（Hawthorne effect）。这种为了获取高质量信号而牺牲生态效度的做法，与 MMLA 的初衷相悖。

这一“生态效度”与“信号质量”之间的根本性矛盾，并非一个无解的困境，而是推动 MMLA 方法论创新的核心动力。它迫使该领域的研究者在两个方向上进行探索：一是开发对噪声更具鲁棒性的单模态特征提取算法；二是通过多模态的交叉验证（cross-validation）来提升整体系统的可靠性。后者尤其体现了 MMLA 的独特优势：当一个模态的信号因噪声而不可靠时（如音频被淹没），可以利用另一模态的信息（如从视频中提取的唇部运动）来进行辅助判断或纠正。^①这种模态间的互补与验证，正是数据融合在解决现实世界数据质量问题时的具体应用，也为下一章深入探讨融合模型提供了实践层面的依据。

为了在突出教育价值的同时保持简洁，原稿中详尽的技术工具列表被整合和精炼。下表聚焦于将核心教育构念与其在 MMLA 中的操作化特征相联系，并强调多模态数据在消解歧义方面的独特能力，而非罗列具体的软件工具。

表 1: 从教育构念到 MMLA 特征的操作化与消歧能力

核心教育构念	操作化维度	MMLA 关键特征（及模态）	推断与消歧能力
学习投入度	认知投入	- 脑电额叶 θ 功率上升（生理） - 话语复杂度（听觉 /NLP）	- 直接测量心智努力。 - 消歧：区分专注的高努力状态与困惑或被动注视。
学习投入度	情感投入	- 面部动作单元（如 AU12- 微笑）（视觉） - 语音韵律（音高、能量）（听觉）	- 评估情绪效价与唤醒度。 - 消歧：阐明“沉默”是源于厌倦（低唤醒）还是挫败（负效价）。
协作学习	互动动态	- 视线同步性（视觉） - 话轮转换频率与均衡度（听觉）	- 衡量联合注意与参与公平性。 - 消歧：区分真正的共享焦点与小组内个体并行工作。
协作学习	知识共建	- 论证性话语（主张、论据、反驳）（听觉 /NLP） - 提问类型（澄清型 vs. 阐释型）（听觉 /NLP）	- 量化话语的认知质量。 - 消歧：超越单纯的发言时长，评估话语对知识建构的贡献。

三、建模与融合架构：从静态融合到动态注意力

在提取了各模态的特征之后，下一个核心步骤是如何将这些异构、多源的信息进行有效融合。数据融合是 MMLA 方法论的核心，其目标是整合来自不同模态的信息，以获得比任何单一模态都更准确、更全面的决策。

（一）LSTM 作为历史基线 vs. Transformer 作为 SOTA 范式

^①参见 M. Worsley et al., “Multimodal Learning Analytics research in the wild: challenges and their potential solutions”, CEUR Workshop Proceedings, Vol-3439, 2023, pp. 45-54.

课堂互动和学习状态本质上是随时间动态演变的过程，因此，能够有效处理时间序列数据的深度学习模型在 MMLA 中扮演着核心角色。

历史基线：循环神经网络（RNN）及其变体：长短期记忆网络（LSTM）通过引入“门控机制”，有效解决了标准 RNN 的梯度消失问题，能够捕捉时间序列中的长期依赖关系。在 MMLA 的早期发展中，一个典型的基线（Baseline）模型架构通常是：并行的单模态编码器（如用于图像的 CNN、用于音频的时频 CNN）分别提取特征，然后将这些特征序列在时间维度上拼接（Concatenation），最后送入一个 LSTM 层进行时序建模和分类。然而，在当前的技术背景下，这种架构已略显陈旧，其顺序计算的特性限制了训练效率和捕捉全局依赖的能力。

当前 SOTA 范式：注意力机制与 Transformer 架构：完全基于自注意力（Self-Attention）机制的 Transformer 模型，摒弃了 RNN 的顺序计算结构，实现了并行化处理，并能更有效地捕捉序列中的全局依赖关系。自其在 NLP 领域取得巨大成功以来，Transformer 架构已被广泛应用于多模态任务，成为构建强大融合模型的骨干网络。一个当前最优（SOTA）的范式通常是：使用强大的单模态预训练编码器（如用于视觉的 Timesformer，用于音频的 Wav2Vec2）提取高质量的特征表示，然后通过堆叠的多层跨模态注意力（Cross-modal Attention）模块进行深度融合，最后送入一个时序解码头（如 Transformer Decoder）进行最终的预测。

将这些技术模型与学习理论相结合，可以获得更深刻的理解。例如，菲利普·温尼（Philip H. Winne）等人提出的自调节学习（Self-Regulated Learning, SRL）模型，将学习视为一个包含计划、监控和反思的循环过程。^①从这个角度看，MMLA 融合模型的目标，本质上是自动捕捉和建模 SRL 过程中的动态监控与调节行为。Transformer 架构能够捕捉长距离的时间依赖关系，这使其特别适合于建模学习者在整个学习任务中（从计划到反思）的策略调整过程，从而为 SRL 理论的实证研究提供了强大的计算工具。^②

（二）核心机制：跨模态注意力与条件依赖性建模

表 2: MMLA 研究的方法学最小基准协议

参数	描述与示例
1. 模态与分辨率	”列出所有使用的数据模态及其技术规格。示例：视觉 (V): 摄像头, 25 fps, 1920x1080; 听觉 (A): 单声道麦克风, 16 kHz; 生理 (Phys): 脑电 (EEG), 32 通道, 256 Hz。”
2. 样本量与时长	描述研究的规模。示例：3 个班级，共 115 名学生；总计 25 小时的有效课堂录制数据。
3. 标注粒度	说明标签所对应的数据时间尺度。示例：片段级（Segment-level），将连续数据切分为 30 秒的片段进行标注；事件级（Event-level），标注特定行为的起止时间。
4. 主要评价指标	定义用于评估模型性能的核心指标。示例：宏平均 F1 分数（Macro F1）、科恩 Kappa 系数（ κ ）、类内相关系数（ICC）、斯皮尔曼等级相关系数（Spearman's ρ ）。
5. 伦理状态与数据治理	阐明伦理审查与数据管理策略。示例：机构审查委员会（IRB）批准；所有参与者均签署知情同意书；原始数据在本地加密存储；数据保留期为 90 天。

从早期、晚期融合到基于注意力的融合，其演进背后反映了对问题本质更深刻的理解。早期和晚期融

^①Philip H. Winne, “Inherent details in self-regulated learning”, 《Educational Psychologist》, 1995, 30(4), pp. 173-187.
^②参见 Jeffrey A. Greene and Roger Azevedo, “A Theoretical Review of Winne and Hadwin’s Model of Self-Regulated Learning: New Perspectives and Directions”, 《Review of Educational Research》, 2007, 77(3), pp. 334-372.

合策略都做出了一个静态的、全局性的假设——即模态间的关系是固定的。然而，在真实的人类交流中，模态间的关系是动态和情境化的。正如第二章所指出的，学生的一句“我不会”（音频）只有在伴随着“皱眉”（视频）时，才更可能表示真实的“困惑”；而伴随着“微笑”，则可能是一种自嘲。语音信号的重要性是根据视频信号提供的上下文而动态变化的。

跨模态注意力（Cross-modal Attention）机制正是为了建模这种“条件依赖性”而设计的。其核心的查询-键-值（Query-Key-Value, QKV）工作原理完美地实现了这一思想。其最小数学公式可以表示为：

在跨模态注意力的场景下，我们可以令查询（Query, Q）来自模态 A 的特征表示，而键（Key, K）和值（Value, V）来自模态 B 的特征序列。这个过程允许一个模态的表示去主动地“探查”并“关注”另一个模态表示中的相关信息。这使得模型能够学会，在解释一个模态时，应该重点关注另一个模态的哪些部分。这不仅是一种技术上的改进，更是一种认知上的飞跃。它让模型能够更忠实地模拟人类多模态交流的协同本质——一个通道中信号的意义常常由另一通道的信号来澄清或调节。因此，这种动态、情境化的融合方法，是解决第二章提出的“多对多”模糊性问题的理想技术方案（表 2）。

（三）一个可复现的方法学基准协议

为了提升 MMLA 研究的可复现性和可比性，本综述提出一个最小化的方法学基准协议，旨在为后续研究提供一个可参考的报告标准。这解决了当前领域内因报告标准不一而导致的研究结果难以比较的困境，通过提供一个清晰的框架，使得其他研究者即便无法获取原始数据，也能复现核心架构并进行有意义的比较（表 2）。

四、“消歧”的可检验评测协议

MMLA 的核心价值在于通过多模态融合来“消歧”。然而，这一论断往往停留在概念层面，缺乏量化的实证支持。为了将“消歧”从一个抽象概念转化为一个可检验、可量化的性能指标，本章提出了一个包含三类可复现实验的评测协议。该协议旨在系统性地评估一个多模态融合模型在面对单模态信息不可靠（有噪声）、不完整（有缺失）或不协调（有错位）时，维持其性能的鲁棒性。这种鲁棒性正是“消歧”能力的直接体现。

（一）协议一：模态鲁棒性曲线

此协议旨在评估模型在某一模态信号质量下降时，利用其他模态信息进行补偿的能力。

任务设置：以一个典型的视听融合任务为例，如通过视频和音频识别人际互动行为。

扰动注入：在测试阶段，保持视觉通道数据不变，向音频通道人为注入不同强度的白噪声。噪声强度可以通过信噪比（Signal-to-Noise Ratio, SNR）来控制，例如，从无噪声（ $\text{SNR} = \infty$ ）逐步降低到强噪声（如 $\text{SNR} = 0 \text{ dB}$ ）。

性能评估：在每个噪声水平下评估模型的性能（如宏 F1 分数）。将不同融合策略（如早期融合、晚期融合、注意力融合）的性能随噪声强度变化的曲线绘制在同一图表中。

预期结果：一个具备良好消歧能力的模型（如跨模态注意力模型）的性能下降曲线应比简单拼接模型（早期融合）更为平缓。当音频信号几乎被噪声淹没时，该模型应能更多地依赖视觉信息，从而维持相对较高的性能。

（二）协议二：脱模态泛化能力

此协议旨在评估模型在训练数据中某一模态完全缺失的情况下，其学习到的跨模态表征的泛化能力。

任务设置：采用“留一法”（Leave-one-modality-out）进行训练。例如，在一个包含视觉（V）、听觉（A）和生理（P）三种模态的数据集上，分别训练三个模型：一个仅使用 V+A 数据，一个使用 V+P 数据，一个使用 A+P 数据。

性能评估：使用完整的 V+A+P 测试集来评估这三个模型的性能。同时，也可以在缺失模态的测试集上进行评估（例如，用 V+A 训练的模型在仅有 V+P 的测试集上表现如何）。

核心指标：关键指标是模型在“训练时未见过的模态组合”上的性能表现，以及与“在全模态数据上训练的模型”之间的性能差距（ $\Delta F1$ ）。一个强大的融合模型应该能够学习到更抽象、更具泛化性的联合表征，从而在面对不完整的模态输入时表现出更小的性能下降。

（三）协议三：时序对齐敏感度

表 3: MMLA 研究的方法学最小基准协议

参数	描述与示例
1. 模态与分辨率	”列出所有使用的数据模态及其技术规格。示例：视觉 (V): 摄像头, 25 fps, 1920x1080; 听觉 (A): 单声道麦克风, 16 kHz; 生理 (Phys): 脑电 (EEG), 32 通道, 256 Hz。”
2. 样本量与时长	描述研究的规模。示例：3 个班级，共 115 名学生；总计 25 小时的有效课堂录制数据。
3. 标注粒度	说明标签所对应的数据时间尺度。示例：片段级 (Segment-level)，将连续数据切分为 30 秒的片段进行标注；事件级 (Event-level)，标注特定行为的起止时间。
4. 主要评价指标	定义用于评估模型性能的核心指标。示例：宏平均 F1 分数 (Macro F1)、科恩 Kappa 系数 (κ)、类内相关系数 (ICC)、斯皮尔曼等级相关系数 (Spearman's ρ)。
5. 伦理状态与数据治理	阐明伦理审查与数据管理策略。示例：机构审查委员会 (IRB) 批准；所有参与者均签署知情同意书；原始数据在本地加密存储；数据保留期为 90 天。

表 5: “消歧”评测协议：任务设置、扰动类型与主指标

协议	任务设置	扰动类型与强度	主指标
模态鲁棒性曲线	视听融合分类	向单一模态（如音频）注入白噪声 / 遮挡 / 串音, 强度从 0% 到 100%（或 SNR 从 ∞ 到 0 dB）	宏 F1 分数随扰动强度的变化曲线
脱模态泛化	留一法训练 (Leave-one-modality-out)	在测试时提供完整的或部分缺失的模态组合	与全模态训练模型相比的性能下降差值 ($\Delta F1$)
时序对齐敏感度	跨模态事件检测	”人为注入模态间的时序错位 ($\pm \Delta t$)，如 80ms, 160ms, 320ms”	性能（如准确率）随 Δt 变化的下降率

此协议旨在评估模型对不同模态数据流之间时间对齐精度的依赖程度，这在真实世界数据同步往往不完美的场景中至关重要。

任务设置：在一个需要精确时序对齐的任务上（如判断话语与对应手势是否一致），人为地在测试数据中注入时间错位。

扰动注入：将一个模态的数据流（如音频）相对于另一个模态（如视频）进行时间上的平移，引入 $\pm \Delta t$ 的延迟或提前（例如， $\Delta t = 80\text{ms}, 160\text{ms}, 320\text{ms} \dots$ ）。

性能评估：报告模型性能（如准确率）随 Δt 增大的下降率。

预期结果：对时序信息不敏感的融合策略（如晚期融合的平均投票）可能对小的错位不敏感，但无法捕捉细粒度的跨模态时序关系。而一个设计精良的注意力模型，其性能应在小的错位下平稳下降，并在错位超过某个阈值后急剧下降，这反映了其对有意义的跨模态时序依赖的建模能力。

通过这三项协议，我们可以将“消歧”这一模糊概念，转化为一系列具体的、可量化的、可跨研究比较的实验基准，从而为 MMLA 模型的评估和迭代提供坚实的证据链。

五、实证支撑：一项基于公开数据集的案例演示

本节为基于公开数据集的二次分析实例。此案例旨在通过一个简洁而具体的实验，为“多模态融合通过消歧提升系统鲁棒性”这一核心论点提供直接的证据支撑。

（一）数据与情境

本案例演示选用 OUC 课堂小组投入度数据集（OUC-CGE）。^① 该数据集包含了大学生在真实课堂协作任务中的多视角视频录像，并提供了小组层面的投入度标注，是研究真实课堂互动动态的宝贵资源。我们选取其视听数据流作为分析对象，任务是预测 30 秒视频片段的小组投入度等级（高 / 中 / 低）。

（二）案例场景：模态鲁棒性曲线的演示

我们模拟一个在真实课堂中极为常见的挑战：音频信道被噪声污染（如邻近小组的讨论声、环境噪音等）。^② 我们旨在通过“模态鲁棒性曲线”协议，量化比较两种不同融合策略的模型在这种情境下的性能表现。

步骤 1（模型构建）：我们构建两个模型。

基线模型（早期融合）：使用预训练的视觉（Timesformer）和听觉（Wav2Vec2）编码器提取特征，在时间维度上简单拼接（concatenate）后，送入一个分类头。

SOTA 模型（跨模态注意力）：使用相同的编码器，但在分类头之前，通过一个跨模态注意力层对视听特征进行深度融合。

步骤 2（扰动注入）：在测试集上，我们保持视觉数据不变，对音频数据人为注入不同强度的高斯白噪声。噪声水平通过信噪比（SNR）来控制，测试点设为 [20, 15, 10, 5, 0, -5] dB，其中 20dB 为轻微噪声，-5dB 为强噪声。

步骤 3（性能评估）：在每个噪声水平下，我们分别评估两个模型的宏 F1 分数，并绘制性能随 SNR 下降的变化曲线。

（三）预期结果与解读

我们期望观察到以下结果：

在低噪声环境（SNR 较高）下，两个模型的性能可能相近，因为此时两个模态的信息都是可靠的。

随着噪声增强（SNR 降低），早期融合模型的性能将急剧下降。这是因为它无法区分信号和噪声，被污染的音频特征会严重“毒化”整个融合后的表征。

相比之下，跨模态注意力模型的性能下降曲线将更为平缓。这正是“消歧”能力的体现。当音频信道变得不可靠时，注意力机制能够学会动态地降低对音频特征的权重（attention weight），转而更多地依赖相对可靠的视觉信道（如学生的身体姿态、视线同步性等）来进行判断。

这个简单的案例演示，将第五章提出的抽象协议转化为一个具体的、可量化的实验结果。它清晰地表明，先进的多模态融合模型（如注意力机制）的价值，并不仅仅在于和平条件下的微小精度提升，更在于其在真实世界复杂、嘈杂环境中维持系统可靠性的鲁棒性。这为本文的核心论点提供了直接的、可复现的证据支撑。当然，必须指出，该曲线的形态会受到样本规模与场景差异的影响（例如，在多语课堂或存在

^①Y. Wang et al., “A Video Dataset for Classroom Group Engagement Recognition”, 《Scientific Data》, 2024, 11(1), 456.

^②M. Worsley et al., “Multimodal Learning Analytics research in the wild: challenges and their potential solutions”, CEUR Workshop Proceedings, Vol-3439, 2023, pp. 45-54.

严重视觉遮挡的教室中,曲线的下降斜率可能会有所不同),这提示我们在外推结论时需保持谨慎。详细的实验参数见附录 C。

六、前沿范式:多模态大语言模型 (MLLMs) 的崛起与应用

在当前人工智能研究的浪潮中,以 GPT-4o 等为代表的多模态大语言模型 (MLLMs) 的出现,正在对 MMLA 领域产生颠覆性的影响。一篇旨在为未来研究提供方向的综述,必须将 MLLMs 从一个简单的“观点陈述”升级为一个可复现、可检验的“应用范式”。MLLMs 正在从根本上重塑 MMLA 的研究流程,它既是强大的赋能工具,也带来了前所未有的挑战。

(一) MLLMs 带来的范式重塑与研究民主化

零样本 / 少样本特征提取: 传统 MMLA 需要为每个特定行为 (如姿态、表情) 训练专门的模型,这需要大量的标注数据和深厚的技术专长。而现在,研究者可以直接利用预训练的 MLLMs,通过精心设计的提示 (prompting),让模型直接从视频或音频中“理解”并提取出复杂的学习行为特征 (如学生的协作姿态、讨论的投入程度),极大地降低了研究门槛和数据标注成本。^① 这预示着 MMLA 研究的潜力可能被“民主化”,使得更多教育研究者能够进行以往只有少数技术专家才能开展的复杂分析。

生成式反馈与脚手架: MLLMs 不仅能“分析”,还能“生成”。未来的 MMLA 系统可以利用其生成能力,为学生提供个性化的解释、总结,甚至创建新的练习题,从而构建更智能、更具适应性的反馈闭环。这种能力正在将 MMLA 的应用从“评估”拓展到“干预”。

统一的端到端建模: MLLMs 为在一个统一的 Transformer 架构中端到端地处理和融合文本、视觉、听觉等多种模态提供了可能,有望替代当前分模块、分阶段的处理与融合流程,简化系统架构。

(二) 一个可复现的 MMLA-MLLMs 课堂评测最小闭环范式

为了将上述潜力转化为可执行的研究流程,本综述提出了一个 MMLA-MLLMs 课堂评测的最小闭环范式。该范式旨在提供一个可复现的草案,明确定义了从任务到风险缓解的全过程,使其他研究者能够在此基础上进行复现、改进和比较。

方法盒: MMLA-MLLMs 课堂评测最小闭环 (可复现草案)

任务: 片段级课堂投入度分级与协作行为标签。

输入: 30 秒视频 (25 fps)、单声道音频 (16 kHz); 两版对照: □ 原始; □ 加噪 (SNR=10/0 dB)。

流程:

传统 MMLA 分支:

单模态编码 (ResNet/Timesformer; Wav2Vec2; 生理信号 1D-CNN)。

跨模态注意力堆叠 (2-4 层, $d_{model} = 512, h = 8$)。

时序解码 (Transformer decoder, $L = 2$)。

MLLM 零 / 少样本分支:

零样本推理: 使用指令提示 (few instructions) 直接对视频片段进行分析。

少样本推理: 通过检索增强生成 (RAG), 向提示中注入 3 条已标注的示例, 引导模型进行更准确的上下文学习。

输出: 标签 + 证据时间戳 / 字幕片段。

指标: 宏 F1、AUPRC (Precision-Recall 曲线下面积)、证据可解释度比率 (输出包含可核查的时间戳 / 字幕片段的比例); 鲁棒曲线 (噪声 / 遮挡 / 错位)。

^①参见 A. Author et al., “A Survey on MLLMs in Education: Application and Future Directions”, 《Journal of Future Tense in Technology Integration》, 2024, 16(12), 467.

超参：学习率 $1e-4$ ，批量大小 32，预热步数 1000 steps，优化器 AdamW，早停耐心值 10。

风险缓解：边缘端 PII 去标识（人脸模糊 / 声纹哈希）、本地推理优先、分组公平性报告。

（三）战略困境：开放 API 的便捷性 vs. 可控模型的严谨性

MLLMs 的崛起将 MMLA 领域推向了一个关键的十字路口。一方面，使用商业闭源 API（如 GPT-4o）提供了无与伦比的便捷性和强大的零样本能力，极大地加速了研究探索的进程。但另一方面，这也带来了严峻的科学与伦理风险。

模型“幻觉”与可靠性问题是首要挑战。MLLMs 可能生成看似合理但与事实不符的分析结果，即所谓“幻觉”（hallucination）。在教育评估这一高风险场景中，基于幻觉的错误判断可能对学生造成实质性伤害。

深层偏见与不透明性同样值得警惕。由于 MLLMs 在海量且未经筛选的互联网数据上进行预训练，它们容易内化并放大社会中原有的偏见。而模型结构与训练数据的不透明，使得外部审计几乎不可能进行，从而对教育公平构成潜在威胁。

数据隐私与主权问题更加复杂。当前主流的强大 MLLMs 多以闭源 API 形式提供服务，这意味着研究机构若要利用其分析能力，往往需要将包含学生面部、声音等高度敏感个人信息的原始数据上传至第三方商业公司。这在需要严格遵守 GDPR、PIPL 等数据保护法规的情况下，带来了严峻的伦理与法律风险。

这种发展趋势迫使研究者做出战略性抉择：是继续依赖商业 MLLMs，享受其强大性能与便利，同时承担不透明和不可控的风险；还是坚持自主构建规模更小、更透明、可本地部署、专为教育场景设计的模型。这不仅是一条技术路线的选择，更是一个关乎科学伦理与社会责任的深层命题。

七、实践中的部署：挑战、治理与在地化

尽管 MMLA 在理论和技术上取得了进展，但将其从实验室环境推广到真实、复杂的课堂中，仍然面临着一系列严峻的技术、伦理和社会挑战。一篇优秀的综述不应止步于问题的罗列，而应深入其内核进行批判性反思，并提出可操作的解决方案。

（一）真实世界部署的技术挑战

在真实课堂这一充满不可控因素的“嘈杂”环境中，背景谈话、桌椅移动、环境光线变化以及学生身体的相互遮挡等，都会导致数据质量下降并引入大量噪声。来自不同传感器、不同采样率的数据流要实现毫秒级的精确对齐，在技术上具有极大的挑战。同时，诸如脑电帽、生理手环等可穿戴设备往往会让师生感到不适或不自然，进而影响其行为的真实性与数据的有效性。

此外，处理和分析海量多模态数据流需要强大的计算资源，这使得 MMLA 技术难以实现大规模、常态化的应用。更为复杂的是，在特定环境中训练出的模型往往缺乏足够的泛化能力，当其被应用到具有不同学生群体、教学风格乃至文化背景的新环境中时，性能常常会出现明显下降。这些问题共同构成了 MMLA 从实验室走向真实教育场景的主要技术障碍。

（二）一个 PIPL GDPR 双栈合规的治理框架

随着 MMLA 技术的深入应用，其引发的伦理问题必须予以高度重视。这不仅仅是一个合规问题，更是一个关乎技术设计本身的架构问题。伦理与合规要求，特别是数据最小化和隐私保护原则，应成为系统设计的内在驱动力，而非事后的附加约束。为此，本综述提出了一个旨在同时满足中国《个人信息保护法》（PIPL）和欧盟《通用数据保护条例》（GDPR）核心要求的“双栈”治理框架。

该框架的核心思想是，伦理原则直接指导技术架构的选择。例如，PIPL 和 GDPR 都强调数据最小化和目的限制原则。PIPL 在同意和跨境数据传输方面有更严格的要求，例如它不像 GDPR 那样提供“合法利益”作为处理个人信息的法律依据，这使得“明确同意”变得尤为重要。为了满足这两个法规中最严格

的要求,系统架构应天然地倾向于一个分布式的、边缘计算优先的设计范式,而非一个中心化的、云端处理一切的范式。

具体而言,该框架可以通过以下操作性清单来落实:其一,最小化收集,仅采集达成教与学目的之必要模式;默认关闭生理通道,需二次明确同意方可开启。其二,在地存储与边缘计算,原始视频/音频等包含生物特征的个人信息(PII)应在教室的边缘计算设备上进行处理和加密存储。云端仅保留和分析去标识化或匿名化后的特征向量,从而从根本上降低数据泄露风险并满足数据本地化要求。其三,可撤回权,学生/家长应能通过简单的界面(如一键操作)随时撤回其同意,并要求删除其个人数据。系统默认数据留存期不应超过为教学改进目的所必需的最短时间(例如,90天)。其四,分级可见性与访问控制,系统应实施严格的访问控制。教师仅能看到班级或小组层面的汇总统计数据和匿名化的行为模式。查看任何与个体学生相关的画像或报告,必须获得学生和家长的二次明确同意。其五,偏见审计与公平性报告,系统开发者和使用者有责任定期进行偏见审计。应按性别、年级、母语/方言等受保护的群体属性,分组报告模型的性能指标(如F1分数)和错误率(如假阳性率),并公开显著的性能差异($\Delta F1, \Delta FPR$)。其六,人机协同的最终解释权,必须明确,MMLA系统仅作为教师的辅助工具,提供提醒、证据展示和数据洞察。所有最终的教育判断、评估和干预决策,其权力和责任都归于人类教师,而非算法。

(三) 全球挑战,在地化对策:一个比较的视角

将MMLA技术应用于不同国家的教育情境,必须充分考虑其独特的“在地化”挑战。这不仅是难题,更是研究能够做出独特贡献的机遇。通过对中西方主流课堂挑战的比较,我们可以更深刻地理解为何一个“放之四海而皆准”的MMLA标准是不可行的,并进而提出一个更具适应性的“全球本土化”(glocal)框架。

中国的课堂环境通常面临着结构性的挑战,如普遍存在的45-60人的大班额,这导致了严重的相互遮挡问题,使得基于单摄像头的个体行为分析变得极为困难。同时,中国丰富的方言对通用ASR模型的准确性构成了巨大挑战。

欧洲/北美地区的教育改革往往更强调个别化学习(individualized learning)和灵活分组,这要求MMLA系统具备更强的动态追踪和个体识别能力。^①此外,日益增多的移民带来了多语课堂(multilingual classrooms)的挑战,即一个小组内可能有学生使用多种不同的语言进行交流,这对基于NLP的分析构成了根本性障碍。

全球的MMLA研究社群,特别是SoLAR,一直致力于推动该领域的标准化,以促进研究的可比性和可复现性。然而,上述比较清晰地揭示了标准化努力与在地化需求之间的内在张力。一个为解决欧洲小班额、多语环境设计的系统,可能完全无法应对中国大班额、多方言的现实。

这种张力并非不可调和。它指向了MMLA未来发展的一个关键方向:构建“全球本土化”(glocal)框架。这意味着我们需要的不是一个僵化的全球标准,而是一个包含全球共通核心原则(如伦理框架、可复现性协议)和可配置的在地化模块的混合框架。该框架应允许研究者和实践者根据本地的结构(班级规模)、语言(方言/多语)和文化(互动规范)现实,对模型的参数、特征集乃至所要测量的构念本身进行适配和微调。例如,一个“glocal”的ASR模块,其底层架构是全球共通的,但它必须提供接口,以便使用本地的方言数据进行快速的域适配。同样,一个“glocal”的投入度模型,其核心维度(行为、情感、认知)是共通的,但其具体的行为指标(如“提问”的频率和形式)必须与本地教师共同定义,以避免第二章所讨论的、因文化资本差异而导致的“符号暴力”。

^①参见 M. Cukurova et al., “The promise and challenges of multimodal learning analytics”, 《British Journal of Educational Technology》, 2020, 51(5), pp. 1479-1484.

表 7: MMLA 在地化挑战的比较分析

维度	典型中国课堂挑战	典型欧洲 / 北美课堂挑战	对“全球本土化” MMLA 设计的启示
结构	大班额（45-60 人）：严重的视觉遮挡，高密度的音频串音。	灵活 / 个别化分组：动态、小规模的小组，更多的移动。	系统架构：前者需要多摄像头融合与先进的音频源分离技术；后者需要鲁棒的个体追踪与重识别算法。
语言	方言多样性：基于普通话训练的 ASR 模型在地方方言上性能下降。	多语环境：一个小组内使用多种不同语言，NLP 分析困难。	ASR/NLP 模型：前者需要利用本地数据进行域适配；后者可能需要转向语言无关的特征（如韵律、话轮）或开发多语言识别模型。
社会文化（布迪厄视角）	以教师为中心的互动规范：学生的沉默可能被视为尊重，而非不投入。	以学生为中心的探究式规范：积极、甚至略显混乱的发言被视为常态。	构念操作化：“投入度”模型必须与本地教师共同设计，定义符合文化情境的行为指标，避免将一种文化规范强加于另一文化。

结论

本文的核心理论与方法论贡献可凝练为以下三点：第一，提出“方法—证据—规范”三维统一评估框架，本文独创性地构建了一个整合技术可行性、实证有效性与伦理社会适切性的统一分析框架，为 MMLA 研究提供了一个超越单纯技术盘点的、更具批判性与整合性的评估与设计准则。第二，给出可复现评测协议并配套开源数据二次分析演示，为将 MMLA 的核心价值“消歧”从抽象概念转化为可检验的指标，本文设计了包含噪声鲁棒性曲线、脱模态泛化与时序对齐敏感度的标准化评测协议，并通过一个基于公开数据集的二次分析案例，具体展示了其可操作性，为提升领域研究的严谨性与可复现性提供了工具。第三，将 CLT 与文化资本并置为 MMLA 的双透镜，并提出“glocal”在地化设计准则，本文创新性地引入认知负荷理论（CLT）与布迪厄的文化资本理论作为跨学科分析透镜，深刻揭示了 MMLA 在追求教学效率与维护教育公平之间的内在张力。在此基础上，通过国际比较，本文进一步提出了超越“标准化”与“在地化”二元对立的“全球本土化”（glocal）设计准则，为 MMLA 在不同文化语境下的适应性发展提供了理论指导。

本文系统地探讨了基于多模态学习分析（MMLA）对课堂协作与投入度进行自动评估的理论基础、关键技术、模型架构及未来挑战。研究独创性地提出了“方法—证据—规范”的三维统一框架，旨在为该领域的科学严谨性与实践落地性提供整合性的支撑。从方法论上看，本文梳理了一个从多模态特征工程到深度融合建模的综合性技术框架。在特征提取层面，本文修正了关于认知负荷与 EEG 信号关系的关键性学术表述，明确了“额叶 θ 功率上升、顶叶 α 功率下降”的主流模式及其边界条件。在模型构建层面，本文清晰地将技术演进脉络定位为从以 LSTM 为代表的历史基线，迈向以跨模态注意力机制为核心的 Transformer 架构这一当前最优范式，并将其与自调节学习（SRL）等教育理论进行了有机关联。在证据链方面，为解决 MMLA 的核心价值——“消歧”——长期停留在概念层面而缺乏量化实证的困境，本文提出了一个由噪声鲁棒性曲线、脱模态泛化和时序对齐敏感度构成的可复现评测协议。并通过一个基于公开数据集的案例演示，具体展示了该协议如何为融合模型的鲁棒性优势提供可量化的实证基础。在规范性层面，本文超越了对伦理问题的简单罗列，提出了一个兼顾中国《个人信息保护法》（PIPL）与欧盟《通用数据保护条例》（GDPR）核心要求的双栈合规治理框架。该框架强调伦理原则对技术架构的驱动作用。此外，本文引入了认知负荷理论与布迪厄的文化资本理论作为跨学科透镜，深刻揭示了 MMLA 在

追求教学效率与维护教育公平之间存在的内在张力。通过对中西方课堂挑战的比较,本文进一步论证了超越“标准化”与“在地化”二元对立的“全球本土化”(glocal)MMLA框架的必要性。特别地,面对多模态大语言模型(MLLMs)这一前沿技术,本文将其从一个颠覆性的“观点”落地为一个可执行的“范式”。通过提出一个包含任务定义、推理路径、评测指标和风险缓解策略的最小闭环框架,本文为研究者利用MLLMs开展可复现、负责任的教育研究提供了操作性指南,并深刻揭示了该领域在拥抱商业闭源API与构建领域专用可控模型之间面临的战略抉择。

综上所述,MMLA的最终愿景并非是为了构建一个无所不知的监控系统来评判师生,而是为了创造一个能更好理解学习、更精准支持学习者、更有力赋能教师的智能环境。通过将人工智能的分析能力与人类教师的教育智慧相结合,MMLA有望在未来教育中扮演“增强智能”的角色,助力实现更加个性化、自适应、公平和人性化的教育理想。未来的研究必须将技术创新与教育理论、伦理规范和人的价值进行深度绑定,大力发展可解释AI以建立信任,构建完整的反馈与干预闭环以彰显教育价值,并开展长期纵贯研究以评估真实影响。

引用本文 涂蓓. 从信号到意义:课堂协作与投入度的多模态学习分析[J]. 人工智能教育研究, 2025, 1(2):30-46. <https://doi.org/10.6914/aiescn.010203>.

Cite This Article Bei TU.(2025). From Signals to Meaning: Multimodal Learning Analytics of Classroom Collaboration and Engagement. *Artificial Intelligence Education Studies(Réngōng zhìnéng jiàoyù yánjiū)*, 1(2):30-46. <https://doi.org/10.6914/aiescn.010203>

From Signals to Meaning: Multimodal Learning Analytics of Classroom Collaboration and Engagement

Bei TU

1.Hong Kong Integration Research Institute 2.Artificial Intelligence Education Studies Press

Abstract This paper reviews advances in Multimodal Learning Analytics (MMLA) for understanding classroom collaboration and engagement. It synthesizes feature engineering across visual, audio, physiological, and interaction data, comparing early, late, and attention-based fusion models, and proposes reproducible evaluation protocols for robustness, cross-modality generalization, and alignment sensitivity. Integrating Cognitive Load Theory and Bourdieu's Cultural Capital, it explores the tension between learning measurement and educational equity, while outlining a PIPL-GDPR dual-stack governance framework and localized design principles. The paper further discusses zero-shot classroom assessment using Multimodal Large Language Models (MLLMs). It offers a holistic paradigm integrating technical feasibility and ethical accountability for explainable, reproducible, and equitable MMLA.

Keywords Multimodal Learning Analytics; Classroom Collaboration; Engagement; Cross-modal Attention; Ethical Governance

(责任编辑:黄欣 cpcl@cpcl.hk)